seit 1558

Friedrich-Schiller-Universität Jena
Fakultät für Sozial- und Verhaltenswissenschaften
Institut für Psychologie

Dissertation

Faire Vergleiche in der Schulleistungsforschung – Methodologische Grundlagen und Anwendung auf Vergleichsarbeiten

Dissertation
zur Erlangung des akademischen Grades
doctor philosophiae (Dr. phil.)

vorgelegt dem Rat der Fakultät für Sozial- und Verhaltenswissenschaften
der Friedrich-Schiller-Universität Jena
von Dipl.-Psych. Christiane Fiege
geboren am 29.08.1981 in Heiligenstadt

Gutachter:

1. Prof. Dr. Rolf Steyer (Friedrich-Schiller-Universität Jena)

2. Prof. Dr. Andreas Frey (Friedrich-Schiller-Universität Jena)

Tag der mündlichen Prüfung: 01. Juli 2013

Für mein Sternchen

Danksagung

Mit der Abgabe meiner Dissertationsschrift möchte ich mich bei allen Personen bedanken, die mich in den letzten Jahren bei der Entstehung dieser Arbeit unterstützt haben.

Mein besonderer Dank gilt meinem Doktorvater Prof. Dr. Rolf Steyer, der es vermochte, mein Interesse für die Kausalitätstheorie zu wecken und mich hierbei maßgeblich förderte. Seine mathematisch-logische Präzision hat mich beim Verfassen dieser Arbeit entscheidend beeinflusst. Prof. Dr. Andreas Frey danke ich für seine Unterstützung, insbesondere auf den letzten Metern dieser Arbeit, für sein fachliches Interesse sowie seine Bereitschaft, diese Arbeit zu begutachten. Bei Prof. Dr. Benjamin Nagengast bedanke ich mich herzlichst für seinen konstruktiven fachlichen Rat und seine Förderung. Dr. Christof Nachtigall gehört mein Dank sowohl in fachlicher als auch in persönlicher Hinsicht. Durch ihn ist die Bearbeitung der empirischen Fragestellung dieser Arbeit erst möglich geworden.

Norman Rose und Axel Mayer haben mich während des gesamten Arbeitsprozesses als Kollegen und als Freunde durch so manches Hoch und Tief begleitet. Ihre methodischen Anregungen, ihre konstruktive Kritik und die zahllosen Gespräche mit ihnen haben mich und diese Arbeit besonders bereichert.

Mein Dank gilt auch meinen ehemaligen Hilfskräften Franziska Lemke, Marie-Ann Sengewald und Anna Zimmermann. Sie haben durch ihre zuverlässige Arbeit zum Gelingen des BMBF-Projektes *Faire Vergleiche* beigetragen, welches die Grundlage dieser Dissertationsschrift darstellt. Einen besonderen Dank möchte ich Katrin Schaller und Marcel Bauer aussprechen. Sie haben mit mir zusammen auf dem zuweilen auch steinigen Weg so manche Hindernisse aus dem Weg geräumt.

Mein Weg zur Dissertation begann am Lehrstuhl für Methodenlehre und Evaluationsforschung an der Friedrich-Schiller-Universität Jena. Meine damaligen Kollegen standen mir in vielen Fragen hilfreich zur Seite und haben mich mit Rat und Tat stets unterstützt. Dies war eine sehr intensive, lehrreiche und oft auch herausfordernde Zeit, die mich nachhaltig beeinflusst hat und an die ich mich stets gern erinnere. Die letz-

ten Meter dieser Arbeit habe ich in Tübingen zurückgelegt – in diesem Zusammenhang danke ich Herrn Prof. Dr. Ulrich Trautwein und den Kollegen der Abteilung Empirische Bildungsforschung und Pädagogische Psychologie an der Eberhard Karls Universität Tübingen, die mir die Möglichkeit zum Abschließen dieser Arbeit gegeben haben.

Meinen Freunden danke ich für ihren Rückhalt und ihre Geduld. Vor allem Claudia Raschdorf danke ich zutiefst für ihre treue Freundschaft, die mir in den letzten Jahren viel Kraft gegeben hat. Auch bei Helen Hertzsch und Christian Breidenstein bedanke ich mich herzlichst. Sie haben mir stets Mut zugesprochen und mich in meiner Arbeit – insbesondere bei den letzten „Besenstrichen“ – bestärkt.

Mein besonderer Dank gilt meiner Familie. Meine Eltern, Brigitte und Jürgen Fiege, haben es mir durch ihr Vertrauen und ihren Glaube in meine Fähigkeiten ermöglicht, diese Arbeit anzufertigen. Meinem Bruder René danke ich für so manche Motivationsarbeit – er gab mir zahlreiche hilfreiche Tipps, die sich auf seinem Weg zur Promotion bewährt hatten.

Schließlich möchte ich Tim Loßnitzer aus tiefsten Herzen danken. Er hat mit viel Ausdauer, Gelassenheit und Humor der „unerträglichen Leichtigkeit des Seins“ ein angemessenes Gewicht verliehen.

Christiane Fiege
Tübingen, Mai 2013

 ausal inference may well be the holy grail of quantitative research in the social sciences, but it should not be proclaimed lightly.

Briggs & Domingue (2011)

Zusammenfassung

Moderne Educational-Accountability-Systeme zeichnen sich insbesondere durch eine zunehmende Evidenzbasierung aus (Ryan & Shepard, 2008). Dabei wird häufig die Leistung der Schüler als zentrales Output-Kriterium zur Evaluation der Leistungsfähigkeit eines Bildungssystems genutzt. Dies trifft – wenn auch erst seit jüngerer Zeit – ebenso auf die Bundesrepublik Deutschland zu. Im Jahr 2006 beschloss die Kultusministerkonferenz die sog. *Gesamtstrategie zum Bildungsmonitoring* (KMK, 2006). Diese umfasst Maßnahmen zur systematischen und wissenschaftlich fundierten Evaluation von Ergebnissen des Bildungssystems, die auf verschiedenen Ebenen des Bildungssystems ansetzen. Einen wichtigen Bestandteil der Gesamtstrategie bilden die landesweiten Vergleichsarbeiten, die den Leistungsstand von Schülern mittels standardisierter und standardbezogener Tests erheben. Ein gemeinsames Ziel dieser Vergleichsarbeiten ist es, durch den Vergleich der Testleistung verschiedener Klassen Aussagen über Unterrichtseffekte zu ermöglichen. Diese sollen Ansatzpunkt für Unterrichts- und Schulentwicklungsmaßnahmen sein. Um zu *fairen Vergleichen* zu gelangen, müssen die unterschiedlichen Ausgangsvoraussetzungen der Schüler – wie sozioökonomischer Status oder Muttersprache – berücksichtigt werden. Deshalb werden statistische Adjustierungsverfahren verwendet, die Unterschiede bezüglich dieser außerschulischen Einflussgrößen des Lernens (sog. Kovariaten) zu berücksichtigen suchen.

Derzeit gibt es im Rahmen von Vergleichsarbeiten verschiedene Adjustierungsverfahren, welche sich hinsichtlich der methodischen Vorgehensweise sowie der Art und Anzahl der dabei berücksichtigten Kovariaten unterscheiden (vgl. Fiege, Reuther & Nachtigall, 2011). Es finden sich starke regionale und institutionelle Unterschiede hinsichtlich der Methodik, d. h. in der Art und Weise wie Kovariaten berücksichtigt werden. Die Palette reicht von wenig theoretisch fundierten Ad-hoc-Verfahren bis hin zu elaborierten, modellbasierten Adjustierungsverfahren. Die Wahl der Methode hat jedoch Einfluss auf die Ergebnisse, so dass sich insbesondere die Frage stellt: Welche Verfahrensweise ist die richtige?

In der vorliegenden Arbeit sollen die derzeit angewendeten Adjustierungsverfahren systematisiert und hinsichtlich verschiedener Kriterien evaluiert werden. Zudem werden diese Verfahren mittels eines Vergleichs mit anderen Educational-Accountability-Systemen in den internationalen Kontext eingeordnet. Zur Beurteilung der Fairness sollen die Adjustierungsverfahren aus kausaltheoretischer Perspektive (Steyer, Partchev, Kröhne, Nagengast & Fiege, 2011) betrachtet werden. Die Interpretierbarkeit der Effektschätzungen einzelner Klassen – d. h. der potenziell fairen Vergleiche – als kausale Effekte des Unterrichts wird diskutiert. Anschließend werden die aus der Theorie ableitbaren Implikationen bezüglich der Kovariaten- und der Modellauswahl dargestellt.

Als Anwendungsbeispiel werden Schulleistungsdaten aus dem Thüringer Projekt *Kompetenztest.de* verwendet. Die Sensitivität der klassenspezifischen Effektschätzungen gegenüber der Modellspezifikation und der Auswahl der Kovariaten wird analysiert, wobei insbesondere die Relevanz der Kovariate Vorwissen betrachtet wird. Leider liegen nur selten längsschnittliche Daten für deutsche Vergleichsarbeiten vor. In der Literatur finden sich jedoch zahlreiche Hinweise, dass der Vortest bzw. das Vorwissen eine der wichtigsten Größen ist, die bei fairen Vergleichen zu berücksichtigen ist (z. B. Steiner, Cook, Shadish & Clark, 2010). Anhand eines Modellvergleichs im Rahmen einer empirischen Reanalyse von Thüringer Kompetenztestdaten wird u. a. aufgezeigt, welchen Einfluss die Hinzunahme des Vorwissens der Schüler als zusätzliche Kovariate auf die Effektschätzungen hat. Ziel ist es, die Bedeutsamkeit des Vortests sowie weiterer Modifikationen des Adjustierungsmodells zu quantifizieren.

Die zentralen Befunde dieser Arbeit sind: (1) Faire, kausal interpretierbare Vergleiche sind theoretisch möglich, im Kontext von Schulleistungsuntersuchungen wie den landesweiten Vergleichsarbeiten jedoch nicht realisierbar. Realistisch sind *fairere* Vergleiche, die als deskriptive Maße im Kontext von Low-Stakes Assessment Systemen informativen Nutzen haben. (2) Es gibt nicht *das* richtige Adjustierungsverfahren. Bei der Modellselektion und der Wahl der Kovariaten sind neben der Fairness auch Praktikabilitätsaspekte zu berücksichtigen. Zudem sollten dabei nicht allein die Varianzaufklärung und entsprechende inferenzstatistische Tests, sondern stets auch die Sensitivität bzw. Stabilität der Effektschätzungen auf Ebene einzelner Klassen als Kriterium herangezogen werden. (3) Entscheidend ist die richtige Auswahl der Kovariaten. Wenn möglich, sollte das fachspezifische Vorwissen in die Berechnung fairerer Vergleiche einbezogen werden. Dies unterstreicht die Bedeutung flächendeckender längsschnittlicher Designs für die weitere Praxis fairerer Vergleiche.

Inhaltsverzeichnis

1 Einführung **1**

1.1 Anliegen der Arbeit 3

1.2 Struktur der Arbeit 4

2 Vergleichsarbeiten: Definition, Ziele und die Bedeutung fairer Vergleiche **5**

2.1 Empirische Bildungsforschung und ihre Gegenstandsbereiche 5

2.2 Von der Input- zur Output-Orientierung 7

2.3 Bildungsstandards 11

2.4 Die Gesamtstrategie der KMK zum Bildungsmonitoring 13

2.5 Vergleichsarbeiten 14

2.5.1 Ziele und Funktionen von Vergleichsarbeiten 16

2.5.2 Vergleichsarbeiten als spezielle Form der Evaluation 19

2.5.3 Bezugsnormen bei der Leistungsbewertung: Womit vergleichen Vergleichsarbeiten? 22

2.5.4 Faire Vergleiche in Vergleichsarbeiten 24

2.6 Zusammenfassung 25

3 Kausale Effekte: Faire Vergleiche und die Theorie kausaler Effekte **27**

3.1 Gegenstandsbestimmung 28

3.2 Terminologie und Grundkonzepte 30

3.2.1 Single-Unit-Trial 32

3.2.2 Kausalitätsraum 36

3.3 Kausale Effekte 40

3.3.1 True-Outcome-Variable und True-Effect-Variable 41

3.3.2 Durchschnittliche und bedingte kausale Effekte 42

3.3.3 Effektparametrisierung kausaler Effekte 46

3.4 Identifikation kausaler Effekte 49
3.4.1 Unverfälschtheit 49
3.4.2 Kausalitätsbedingungen 50
3.5 Kausale Effekte und faire Vergleiche in Vergleichsarbeiten 53
3.5.1 Der intendierte kausale Effekt: Definition und Identifikation . . 54
3.5.2 Der adjustierte Effekt $E(\delta_{adj} \mid X{=}x)$ in Vergleichsarbeiten und seine kausaltheoretische Verortung: Kausal oder nicht kausal, das ist hier die Frage 57
3.5.3 Eigenschaften der adjustierten Effektfunktion $E(\delta_{adj} \mid X)$ 61
3.5.4 Konsequenzen der Effektdefinition für die Interpretation 65
3.6 Zusammenfassung 67

4 Adjustierungsmodelle: Die Berechnung fairer(er) Vergleiche 69
4.1 Systematisierung statistischer Adjustierungsverfahren im Kontext von Vergleichsarbeiten 69
4.1.1 Methodisches Vorgehen: Eine Quellenanalyse 70
4.1.2 Kategorien von Adjustierungsstrategien 70
4.1.3 Kriterien zur Bewertung der Adjustierungsstrategien 78
4.1.4 Adjustierungsstrategien in den Bundesländern 81
4.2 Einordnung in den internationalen Kontext: Ein Vergleich mit den USA und England 86
4.2.1 State Achievement Tests in den USA 87
4.2.2 Key Stage Tests in England 92
4.2.3 USA, England und Deutschland im Vergleich 100
4.3 Zusammenfassung 104

5 Fragestellungen 106
5.1 Faire Vergleiche als kausale Unterrichtseffekte 106
5.2 Zwei zentrale Facetten fairer(er) Vergleiche 108
5.3 Aggregation bisheriger Befunde aus der Perspektive verschiedener methodischer Zugänge 109
5.3.1 Der analytische Zugang 110
5.3.2 Der simulationsbasierte Zugang 112
5.3.3 Der empirische Zugang 114

5.4 Fragestellungen und Hypothesen . . . 122
5.4.1 Offene Fragen bei Vergleichsarbeiten . . . 122
5.4.2 Methodischer Zugang: Empirische Reanalyse von Daten aus Vergleichsarbeiten . . . 123
5.4.3 Hypothesen . . . 124

6 Methode: Empirischer Vergleich verschiedener Adjustierungsmodelle 127
6.1 Die Thüringer Kompetenztests und das Projekt *Kompetenztest.de* . . . 127
6.2 Erhebungsinstrumente und Variablen . . . 130
6.2.1 Kompetenztests in den Fachbereichen Mathematik und Deutsch 130
6.2.2 Kovariaten . . . 134
6.3 Statistische Methoden . . . 137
6.3.1 Design des Modellvergleichs . . . 137
6.3.2 Umgang mit fehlenden Werten . . . 149
6.4 Zusammenfassung . . . 154

7 Ergebnisse: Empirische Befunde aus dem Modellvergleich 155
7.1 Deskriptive Analysen . . . 155
7.1.1 Deskriptive Statistiken . . . 156
7.1.2 Struktur fehlender Werte . . . 160
7.2 Multiple Imputation fehlender Werte . . . 166
7.2.1 Multiple Imputation mit MICE . . . 166
7.2.2 Diagnostik der multiplen Imputation . . . 168
7.3 Modellvergleich . . . 171
7.3.1 Caterpillar-Plots . . . 172
7.3.2 Determinationskoeffizient $R^2_{Y|\mathbf{Z}}$ . . . 181
7.3.3 Korrelationen . . . 198
7.3.4 Change-Plots . . . 206
7.3.5 Transitionsmatrizen . . . 230
7.4 Zusammenfassung . . . 245

8 Diskussion 246
8.1 Faire Vergleiche und kausale Effekte . . . 246
8.2 Faire(re) Vergleiche und statistische Adjustierungsmodelle . . . 248
8.2.1 Systematik statistischer Adjustierungsverfahren . . . 249

8.2.2 Facetten fairer(er) Vergleiche 251
8.2.3 Ergebnisse des Modellvergleichs 253
8.3 Grenzen und kritische Reflexion 261
8.3.1 Weiterer Forschungsbedarf 262
8.3.2 Generalisierbarkeit der Ergebnisse 263
8.4 Conclusio und Ausblick 265
8.4.1 Ausblick 266
8.4.2 Mindestanforderungen für Low-Stakes Assessment 268

Literatur **270**

Anhang **293**

A Abkürzungsverzeichnis **293**

B Vergleichsarbeiten und weitere Formen der Evaluation **296**

C Contextual Models **299**

D Struktur fehlender Werte **301**

E Standardfehler der Effektschätzungen **305**

F Sensitivität der adjustierten Effektschätzungen: Supplement **309**

Abbildungsverzeichnis

2.1 Basismodell der Funktionsweise von Bildungssystemen (in Anlehnung an Scheerens, 2008) 7
2.2 Essentielle Komponenten des Evaluationsprozesses im Kontext von Vergleichsarbeiten (in Anlehnung an Fiege et al., 2011) 21

3.1 Schematische Darstellung verschiedener Ebenen der empirischen Kausalforschung 29
3.2 Venn-Diagramm der Filtration $(\mathfrak{F}_t)_{t\in T}$ mit vier σ-Algebren ($T = \{1, 2, 3, 4\}$) 38
3.3 Minimalbeispiel mit lediglich zwei Treatment-Stufen $X = a$ und $X = b$. Links: $(X{=}x)$-bedingte Erwartungswerte $E(\tau_x\,|X{=}x)$ der True-Outcome-Variablen τ_x und der jeweilige kausale Referenzwert Ref_{causal}. Rechts: $(X{=}x)$-bedingte kausale Effekte $CCE_{x;\,X=x}$ von x und deren Erwartungswerte. Es gelte jeweils $P(X{=}x) = J^{-1}$ für jeden Wert x von X. 63

6.1 Zuständige Institutionen im Rahmen der Testentwicklung, Durchführung und Auswertung der Thüringer Kompetenztests 128
6.2 Kompetenzbereiche in den Bildungsstandards für den Fachbereich Deutsch in der Sekundarstufe I (in Anlehnung an KMK, 2004b) 133
6.3 Rückmeldeformat im Projekt *Kompetenztest.de* (aus Nachtigall et al., 2010). Links: Grafische Darstellung des Klassenmittelwertes und des korrigierten Landesmittelwertes am Beispiel der Mathematikleistung in Klassenstufe 6 (MK6). Rechts: Deskriptive Kennwerte. 139

7.1 Missing-Struktur. Links: Balkendiagramm mit dem Anteil fehlender Werte pro Variable. Rechts: *Aggregation plot* mit allen beobachteten Kombinationen fehlender und beobachteter Werte. Beobachtete Werte sind in Blau, fehlende Werte in Rot dargestellt. 161

7.2 Der weiße Boxplot (links) zeigt die Verteilung der beobachteten Mathematikleistungsscores in Klassenstufe 8 (MK8). Rechts daneben: Parallele Boxplots für die Verteilung von MK8 in Abhängigkeit von der Missing-Struktur bezüglich der Variablen DK8, MK6, DK6, MK3, DK3L, DK3S, BLSF.D, SES.M, SES.D und MUSPR. Die Verteilung von MK8 wird hier in je zwei Gruppen dargestellt; getrennt nach dem Fehlen (rot = *missing*) und Nicht-Fehlen (blau = *observed*) auf den anderen Variablen des Datensatzes. Unterhalb der Boxplots sind die absoluten Häufigkeiten der beobachteten bzw. fehlenden Werte abgetragen. 163
7.3 Nonparametrische Dichteschätzungen aufgrund der beobachteten und imputierten Werte der Mathematikleistung in Klassenstufe 3, 6 und 8 (MK3, MK6 und MK8). 169
7.4 Nonparametrische Dichteschätzungen aufgrund der beobachteten und imputierten Werte der Deutschleistung in Klassenstufe 3, 6 und 8 (DK3L, DK3S, DK6 und DK8). 170
7.5 Caterpillar-Plots der CAM im Fach Mathematik (MK8). Links: Saturierte Parametrisierung (Modell 1). Rechts: Lineare Parametrisierung ohne Interaktionen (Modell 8). 173
7.6 Caterpillar-Plots der VAM im Fach Mathematik (MK8). Links: Bedingt lineare Parametrisierung mit Interaktionen (Modelle 2, 3, 4). Rechts: Lineare Parametrisierung ohne Interaktionen (Modelle 9, 10, 11). 174
7.7 Caterpillar-Plots der CVA im Fach Mathematik (MK8). Links: Bedingt lineare Parametrisierung mit Interaktionen (Modelle 5, 6, 7). Rechts: Lineare Parametrisierung ohne Interaktionen (Modelle 12, 13, 14). 175
7.8 Caterpillar-Plots der CAM im Fach Deutsch (DK8). Links: Saturierte Parametrisierung (Modell 1). Rechts: Lineare Parametrisierung ohne Interaktionen (Modell 8). 177
7.9 Caterpillar-Plots der VAM im Fach Deutsch (DK8). Links: Bedingt lineare Parametrisierung mit Interaktionen (Modelle 2, 3, 4). Rechts: Lineare Parametrisierung ohne Interaktionen (Modelle 9, 10, 11). 178
7.10 Caterpillar-Plots der CVA im Fach Deutsch (DK8). Links: Bedingt lineare Parametrisierung mit Interaktionen (Modelle 5, 6, 7). Rechts: Lineare Parametrisierung ohne Interaktionen (Modelle 12, 13, 14). 179

7.11 Prozentsatz erklärter Varianz an der Gesamtvarianz der Mathematikleistung in Klassenstufe 8 (MK8) 184
7.12 Prozentsatz erklärter Varianz an der Gesamtvarianz der Deutschleistung in Klassenstufe 8 (DK8) 192
7.13 Korrelationen der adjustierten klassenspezifischen Effektschätzungen zwischen den Modellen im Fachbereich Mathematik (MK8) 200
7.14 Korrelationen der adjustierten klassenspezifischen Effektschätzungen zwischen den Modellen im Fachbereich Deutsch (DK8) 204
7.15 Zwei Beispiele für Change-Plots. Links: Vergleich von zwei Modellen (A vs. B), deren Effektschätzungen identisch sind. Rechts: Vergleich zweier Modelle (A vs. C) mit starken Unterschieden hinsichtlich der resultierenden Effektschätzungen. 207
7.16 Change-Plots im Fach Mathematik (MK8): CAM *versus* VAM (saturierte und bedingt lineare Parametrisierung mit Interaktionen) 210
7.17 Change-Plots im Fach Mathematik (MK8): VAM *versus* CVA (bedingt lineare Parametrisierung mit Interaktionen) 210
7.18 Change-Plots im Fach Mathematik (MK8): Modelle mit *versus* ohne MK3 (bedingt lineare Parametrisierung mit Interaktionen) 212
7.19 Change-Plots im Fach Mathematik (MK8): CAM *versus* VAM (lineare Parametrisierung ohne Interaktionen) 214
7.20 Change-Plots im Fach Mathematik (MK8): VAM *versus* CVA (lineare Parametrisierung ohne Interaktionen) 214
7.21 Change-Plots im Fach Mathematik (MK8): Modelle mit *versus* ohne MK3 (lineare Parametrisierung ohne Interaktionen) 215
7.22 Change-Plots im Fach Mathematik (MK8): Bedingt lineare *versus* lineare Parametrisierung 218
7.23 Change-Plots im Fach Mathematik (MK8): VAM mit bedingt linearer *versus* linearer Parametrisierung 218
7.24 Change-Plots im Fach Mathematik (MK8): CVA mit bedingt linearer *versus* linearer Parametrisierung 218
7.25 Change-Plots im Fach Deutsch (DK8): CAM *versus* VAM (saturierte und bedingt lineare Parametrisierung mit Interaktionen) 221
7.26 Change-Plots im Fach Deutsch (DK8): VAM *versus* CVA (bedingt lineare Parametrisierung mit Interaktionen) 221

7.27 Change-Plots im Fach Deutsch (DK8): Modelle mit *versus* ohne DK3 (bedingt lineare Parametrisierung mit Interaktionen) 222
7.28 Change-Plots im Fach Deutsch (DK8): CAM *versus* VAM (lineare Parametrisierung ohne Interaktionen) 224
7.29 Change-Plots im Fach Deutsch (DK8): VAM *versus* CVA (lineare Parametrisierung ohne Interaktionen) 224
7.30 Change-Plots im Fach Deutsch (DK8): Modelle mit *versus* ohne DK3 (lineare Parametrisierung ohne Interaktionen) 226
7.31 Change-Plots im Fach Deutsch (DK8): Bedingt lineare *versus* lineare Parametrisierung 227
7.32 Change-Plots im Fach Deutsch (DK8): VAM mit bedingt linearer *versus* linearer Parametrisierung 228
7.33 Change-Plots im Fach Deutsch (DK8): CVA mit bedingt linearer *versus* linearer Parametrisierung 228

D.1 Der weiße Boxplot (links) zeigt die Verteilung der beobachteten Deutschleistungsscores in Klassenstufe 8 (DK8). Rechts daneben: Parallele Boxplots für die Verteilung von DK8 in Abhängigkeit von der Missing-Struktur bezüglich der Variablen MK8, DK6, MK6, DK3L, DK3S, MK3, BLSF.M, WDH, SES.M, SES.D und MUSPR. Die Verteilung von DK8 wird hier in je zwei Gruppen dargestellt; getrennt nach dem Fehlen (rot = *missing*) und Nicht-Fehlen (blau = *observed*) auf den anderen Variablen des Datensatzes. Unterhalb der Boxplots sind die absoluten Häufigkeiten der beobachteten bzw. fehlenden Werte abgetragen. 302

Tabellenverzeichnis

2.1 Funktionen von Vergleichsarbeiten differenziert nach den Ebenen des Bildungssystems . . . 17
2.2 Vergleich verschiedener Evaluationsformen im Schulkontext (in Anlehnung an van Ackeren und Klemm, 2009) . . . 19

3.1 Übersicht der wichtigsten kausalen Effekte in der Differenz- und Effektparametrisierung . . . 48
3.2 Übersicht der Definitionen von Unverfälschtheit . . . 51
3.3 Vier ausgewählte Kausalitätsbedingungen, die jeweils die $\boldsymbol{Z}$-bedingte Unverfälschtheit implizieren . . . 54

4.1 Kategorien von Adjustierungsstrategien in deutschen Vergleichsarbeiten . . . 72
4.2 Adjustierungsstrategien im Kontext von Vergleichsarbeiten der 3. Jahrgangsstufe (VERA 3) in den einzelnen Bundesländern . . . 82
4.3 Adjustierungsstrategien im Kontext von Vergleichsarbeiten der 8. Jahrgangsstufe (VERA 8) in den einzelnen Bundesländern . . . 83
4.4 Adjustierungsmodelle: USA, England und Deutschland im Vergleich . . . 101

5.1 Wissenschaftliche Studien zu den beiden Facetten fairer(er) Vergleiche vor dem Hintergrund verschiedener methodischer Zugänge . . . 111

6.1 Erhebungszeitpunkte der Kohorte 2004/2005 mit drei Erhebungswellen im Thüringer Schülerlängsschnitt . . . 130
6.2 Die drei Dimensionen des Kompetenzmodells im Fach Mathematik . . . 131
6.3 Übersicht der Variablen . . . 135
6.4 Modellvergleich . . . 147

7.1 Deskriptive Statistiken . . . 157

7.2 Mittelwerte, t-Test und Effektstärken mit der Mathematikleistung in Klassenstufe 8 (MK8) als abhängige Variable und den Indikatorvariablen für fehlende Werte (Response-Indikatoren) als unabhängige Variablen 165
7.3 Varianz der Effektschätzungen $s^2(\bar{\delta}_{adj;x})$ pro Modell im Fach Mathematik (MK8) 176
7.4 Varianz der Effektschätzungen $s^2(\bar{\delta}_{adj;x})$ pro Modell im Fach Deutsch (DK8) 180
7.5 Determinationskoeffizient $R^2_{Y|\mathbf{Z}}$ pro Modell im Fach Mathematik (MK8) . . 183
7.6 R^2-Differenzen genesteter Modelle: Modifikation der Kovariatenselektion im Fach Mathematik (MK8) 186
7.7 R^2-Differenzen genesteter Modelle: Modifikation der Parametrisierung im Fach Mathematik (MK8) 189
7.8 Determinationskoeffizient $R^2_{Y|\mathbf{Z}}$ pro Modell im Fach Deutsch (DK8) 191
7.9 R^2-Differenzen genesteter Modelle: Modifikation der Kovariatenselektion im Fach Deutsch (DK8) 194
7.10 R^2-Differenzen genesteter Modelle: Modifikation der Parametrisierung im Fach Deutsch (DK8) 196
7.11 Veränderungen des Quintil-Rankings der adjustierten Effektschätzungen im Fach Mathematik (MK8) infolge der zusätzlichen Berücksichtigung des fachspezifischen Vorwissens: CAM *versus* VAM 232
7.12 Veränderungen des Quintil-Rankings der adjustierten Effektschätzungen im Fach Mathematik (MK8) infolge der zusätzlichen Berücksichtigung von Kompositionsmerkmalen: VAM *versus* CVA 234
7.13 Veränderungen des Quintil-Rankings der adjustierten Effektschätzungen im Fach Mathematik (MK8) bei bedingt linearer Parametrisierung: Bedingte Unabhängigkeit 235
7.14 Veränderungen des Quintil-Rankings der adjustierten Effektschätzungen im Fach Mathematik (MK8): Bedingt lineare vs. lineare Parametrisierung . . . 237
7.15 Veränderungen des Quintil-Rankings der adjustierten Effektschätzungen im Fach Deutsch (DK8) infolge der zusätzlichen Berücksichtigung des fachspezifischen Vorwissens: CAM *versus* VAM 239
7.16 Veränderungen des Quintil-Rankings der adjustierten Effektschätzungen im Fach Deutsch (DK8) infolge der zusätzlichen Berücksichtigung von Kompositionsmerkmalen: VAM *versus* CVA 240

7.17 Veränderungen des Quintil-Rankings der adjustierten Effektschätzungen im Fach Deutsch (DK8) bei bedingt linearer Parametrisierung: Bedingte Unabhängigkeit 241
7.18 Veränderungen des Quintil-Rankings der adjustierten Effektschätzungen im Fach Deutsch (DK8): Bedingt lineare vs. lineare Parametrisierung 244

B.1 Klassifikation von Schulleistungsuntersuchungen und anderen Evaluationsformen im Bildungskontext 297
B.2 Klassifikation von Schulleistungsuntersuchungen und anderen Evaluationsformen im Bildungskontext (Fortsetzung) 298

D.1 Mittelwerte, t-Test und Effektstärken mit der Deutschleistung in Klassenstufe 8 (DK8) als abhängige Variable und den Indikatorvariablen für fehlende Werte (Response-Indikatoren) als unabhängige Variablen 303

F.1 Veränderungen des Quintil-Rankings der adjustierten Effektschätzungen im Fach Mathematik (MK8) infolge der zusätzlichen Berücksichtigung des fachspezifischen Vorwissens: CAM *versus* VAM 310
F.2 Veränderungen des Quintil-Rankings der adjustierten Effektschätzungen im Fach Mathematik (MK8) infolge der zusätzlichen Berücksichtigung der leistungsmäßigen Klassenkomposition: VAM *versus* CVA 311
F.3 Veränderungen des Quintil-Rankings der adjustierten Effektschätzungen im Fach Mathematik (MK8): Bedingt lineare *versus* lineare Parametrisierung . 312
F.4 Veränderungen des Quintil-Rankings der adjustierten Effektschätzungen im Fach Deutsch (DK8) infolge der zusätzlichen Berücksichtigung des fachspezifischen Vorwissens: CAM *versus* VAM 313
F.5 Veränderungen des Quintil-Rankings der adjustierten Effektschätzungen im Fach Deutsch (DK8) infolge der zusätzlichen Berücksichtigung der leistungsmäßigen Klassenkomposition: VAM *versus* CVA 314
F.6 Veränderungen des Quintil-Rankings der adjustierten Effektschätzungen im Fach Deutsch (DK8): Bedingt lineare *versus* lineare Parametrisierung . . . 315

The task of causal modeling [may be viewed] as an induction game that scientists play against Nature.

JUDEA PEARL (2000)

1 Einführung

Die empirische Schulleistungsforschung hat in Deutschland vor allem in der letzten Dekade stark an Bedeutung gewonnen – waren doch noch Ende des 20. Jahrhunderts Ergebnisse aus nationalen und internationalen empirischen Schulleistungsuntersuchungen eher rar und von Bildungspolitik und Öffentlichkeit wenig beachtet. Insbesondere seit den unbefriedigenden Ergebnissen der PISA-Studie von 2000 (Deutsches PISA-Konsortium, 2001; OECD, 2001), bei der Deutschland im Vergleich zu den anderen teilnehmenden Staaten in allen Fächern nur unterdurchschnittliche Leistungen erreichte, stieg das öffentliche Interesse an Qualität und Effektivität im Bildungswesen. Vor allem die Leistungen der Schüler[1], d. h. der Output schulischer Arbeit, stehen seither verstärkt im Fokus der Öffentlichkeit.

Zur Beurteilung und Sicherung der Qualität schulischer Arbeit beschloss die Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland die sog. *Gesamtstrategie zum Bildungsmonitoring* (KMK, 2006). Diese beinhaltet Maßnahmen zur systematischen und wissenschaftlich fundierten Evaluation von Ergebnissen des Bildungssystems, die auf verschiedenen Ebenen des Bildungssystems ansetzen. Dieses *Educational-Accountability-System* umfasst – neben der Teilnahme an internationalen Schulleistungsuntersuchungen, einer gemeinsamen Bildungsberichterstattung von Bund und Ländern, der zentralen Überprüfung des Erreichens der Bildungsstandards im Ländervergleich – auch die verstetigte Durchführung landesweiter Vergleichsarbeiten in allen Bundesländern.

Die Testergebnisse der Schüler in den Vergleichsarbeiten sollen die Lehrer über den Leistungsstand der Klasse informieren, Stärken und Schwächen in den Leistungen aufzeigen und Grundlage für die Erarbeitung von Maßnahmen zur Unterrichts- und Qualitätsentwicklung sein. Ein häufig nicht explizit formuliertes Ziel landesweiter Ver-

[1] Aus Gründen der besseren Lesbarkeit wird im nachfolgenden Text das generische Maskulinum anstelle der parallelen Nennung von weiblichen und männlichen Wortformen (z. B. Schülerinnen und Schüler) verwendet.

gleichsarbeiten ist die Beurteilung von Unterrichtseffekten auf Ebene einzelner Schulklassen anhand dieser Output-Daten: Evaluiert wird das Ergebnis von Lehre und Unterricht mittels Schülerleistungen. Ergebnisse aus Vergleichsarbeiten sollen demnach Aussagen über die Wirksamkeit des Unterrichts und die Leistungsfähigkeit der Lehrer ermöglichen.

Um die Effekte des Unterrichts beurteilen zu können, werden die Testergebnisse der einzelnen Klassen – bspw. die Klassenmittelwerte der Testleistungen – mit den Ergebnissen anderer Klassen und Schulen verglichen. Ein Problem bei solchen Vergleichen ist, dass Klassenunterschiede nicht nur aufgrund der Unterrichtseffekte zustande kommen können, sondern auch aufgrund unterschiedlicher Ausgangsvoraussetzungen der Schüler – wie bspw. ihr sozioökonomischer Status, ihre Muttersprache oder ihr Geschlecht. Solche Merkmale, die sich dem direkten pädagogischen Einfluss der Lehrkräfte entziehen, können gleichfalls die Leistung der Schüler beeinflussen. Deshalb werden einfache Mittelwertsvergleiche der Testleistungen verschiedener Klassen als unfair angesehen. Um diesen Unterschieden Rechnung zu tragen und somit zu fairen, kausal interpretierbaren Vergleichen zu gelangen, müssen die unterschiedlichen Ausgangsvoraussetzungen der Schüler berücksichtigt werden. Deshalb werden bei der Datenanalyse der Schulleistungsdaten statistische Adjustierungsverfahren verwendet, die Unterschiede bezüglich dieser außerschulischen Einflussgrößen des Lernens – sog. Kovariaten – zu berücksichtigen suchen.

Wie wird in der Praxis mit dieser Problematik umgegangen? Ein Blick in die einzelnen Bundesländer offenbart eine deutliche Heterogenität hinsichtlich der Datenanalyse und Verwendung von Adjustierungsverfahren im Kontext von Vergleichsarbeiten, wobei teilweise auch unadjustierte Vergleiche zurückgemeldet werden. Werden Adjustierungsverfahren zur Erstellung fairer Vergleiche angewandt, so bestehen Unterschiede in der methodischen Vorgehensweise sowie der Art und Anzahl der dabei berücksichtigten Kovariaten.

Doch sind diese fairen Vergleiche als kausale Unterrichtseffekte interpretierbar? Lassen sich solche fairen, d. h. kausal interpretierbaren Vergleiche vor dem Hintergrund einer Theorie kausaler Effekte rechtfertigen? Und welches Adjustierungsverfahren ist das richtige?

1.1 Anliegen der Arbeit

Die vorliegende Arbeit behandelt die Problematik fairer Vergleiche im Kontext von Schulleistungsuntersuchungen mit besonderem Fokus auf landesweite Vergleichsarbeiten in Deutschland. Die Arbeit soll in dreierlei Hinsicht zum Verständnis fairer Vergleiche im Kontext von Schulleistungsuntersuchungen beitragen, um einen verantwortungsvollen Umgang mit Ergebnissen aus Vergleichsarbeiten zu ermöglichen und zu unterstützen.

(1) Erstens sollen die methodologischen Grundlagen fairer Vergleiche analysiert werden. Der Fairness-Begriff wird dabei mittels der Theorie kausaler Effekte präzisiert. Zudem soll vor diesem Hintergrund geprüft werden, ob sich die Interpretation der aus Adjustierungsverfahren resultierenden Mittelwertsvergleiche als kausale Unterrichtseffekte rechtfertigen lässt.

(2) Zweitens sollen die derzeit verwendeten Adjustierungsstrategien zur Berechnung fairer Vergleiche systematisiert werden. Ferner soll geprüft werden, ob sich diese in den internationalen Kontext einordnen lassen und so ggf. empirische Evidenz aus anderen Educational-Accountability-Systemen auch für Adjustierungsverfahren in Vergleichsarbeiten nutzbar ist.

(3) Drittens soll eine empirische Reanalyse von Daten aus Vergleichsarbeiten die Bedeutung der Modellspezifikation und der Wahl der Kovariaten für die Ergebnisse aus fairen Vergleichen aufzeigen. Es soll geprüft werden, ob es möglich und sinnvoll ist, allgemein verbindliche Standards bzw. Richtlinien zum Umgang mit Adjustierungsverfahren in Vergleichsarbeiten zu formulieren.

Diese Arbeit entstand im Rahmen des gleichnamigen Forschungsprojektes „Faire Vergleiche in der Schulleistungsforschung – Methodologische Grundlagen und Anwendung auf Vergleichsarbeiten“ (Projekt *Faire Vergleiche*) am Lehrstuhl für Methodenlehre und Evaluationsforschung der Friedrich-Schiller-Universität Jena. Dieses Projekt wurde vom Bundesministerium für Bildung und Forschung (BMBF) gemäß dem Rahmenprogramm zur Förderung der empirischen Bildungsforschung finanziert[2].

[2]Projekt *Faire Vergleiche*, Förderkennzeichen: 01 GJ 0852, Laufzeit: 01.02.2009 – 31.07.2011, `http://www.fair.uni-jena.de`

1.2 Struktur der Arbeit

Die vorliegende Dissertation ist in einen theoretischen und einen empirischen Abschnitt unterteilt.

Der theoretische Teil umfasst Kapitel 2 bis Kapitel 5. In Kapitel 2 wird zunächst die Entwicklung und Bedeutung einer zunehmend evidenzbasierten Bildungssteuerung skizziert. Vor diesem Hintergrund wird die Funktion von Vergleichsarbeiten im Kontext empirischer Qualitätsentwicklung und -sicherung im Bildungswesen sowie die Bedeutung fairer Vergleiche herausgearbeitet. Dabei wird aufgezeigt, dass faire Vergleiche – implizit oder explizit – mit einer kausalen Konnotation einhergehen. Doch lassen sich solche fairen, d. h. kausal interpretierbaren Vergleiche vor dem Hintergrund einer Theorie kausaler Effekte rechtfertigen? Um dies zu prüfen, wird in Kapitel 3 die allgemeine stochastische Theorie kausaler Effekte nach Steyer et al. (2011) vorgestellt. Die Kausalitätstheorie dient als theoretisches Fundament, um die Bedingungen und die Grenzen kausaler Inferenz transparent zu machen. Die im Rahmen fairer Vergleiche betrachteten Mittelwertsvergleiche werden vor diesem Hintergrund hinsichtlich ihrer kausalen Interpretierbarkeit untersucht. Des Weiteren werden in Kapitel 4 die im Rahmen von Vergleichsarbeiten derzeit verwendeten statistischen Adjustierungsverfahren zur Berechnung fairer Vergleiche systematisiert und in den internationalen Kontext eingeordnet. In Kapitel 5 werden die zentralen Forschungsfragen dieser Arbeit dargestellt. Schließlich werden die Hypothesen spezifiziert, die im zweiten Teil dieser Arbeit empirisch geprüft werden sollen.

Der empirische Teil dieser Arbeit umfasst die Kapitel 6 und 7. Zur Prüfung der postulierten Hypothesen wird im Rahmen einer empirischen Reanalyse von Schulleistungsdaten aus dem Projekt *Kompetenztest.de* ein Modellvergleich verschiedener Adjustierungsmodelle durchgeführt. Im Fokus der Analysen steht die Sensitivität der Effektschätzungen gegenüber Modifikationen der Kovariaten- und Modellselektion. Die Daten entstammen den Kompetenztests – den Vergleichsarbeiten des Freistaates Thüringen. Die verwendeten Erhebungsinstrumente und Variablen werden in Kapitel 6 vorgestellt. Im Zentrum dieses Kapitels steht das methodische Vorgehen und das Design des Modellvergleichs. In Kapitel 7 werden – neben den deskriptiven Analysen und der Analyse der Struktur fehlender Werte – die Ergebnisse des Modellvergleichs dargelegt.

Abschließend werden die Ergebnisse des theoretischen und empirischen Abschnitts in Kapitel 8 zusammenfassend diskutiert.

Denken heißt Vergleichen!

WALTHER RATHENAU (1867 – 1922)

2 Vergleichsarbeiten: Definition, Ziele und die Bedeutung fairer Vergleiche

Bildungspolitik und Bildungsforschung gehen in zunehmendem Maße Hand in Hand: Bildungspolitische Entscheidungen werden häufig maßgeblich durch wissenschaftliche Befunde aus der empirischen Bildungsforschung mitbestimmt. Maßnahmen zur Qualitätssicherung und -entwicklung im deutschen Bildungssystem sollen auf empirischen Befunden beruhen, die wissenschaftlichen Standards genügen. Solche Maßnahmen finden nicht nur auf Bundesebene zum Systemmonitoring statt. Auch auf Ebene der einzelnen Bundesländer werden spezifische Maßnahmen zur Sicherung und Weiterentwicklung der Qualität von Unterricht und Schule durchgeführt. Hierzu zählen auch Vergleichsarbeiten. Eine derartige *evidenzbasierte Bildungssteuerung* hat es jedoch nicht immer gegeben und ist historisch betrachtet – v. a. in Deutschland – eine neuere Entwicklung (vgl. z. B. Bundesministerium für Bildung und Forschung, 2008; Maier, 2009).

2.1 Empirische Bildungsforschung und ihre Gegenstandsbereiche

Die empirische Bildungsforschung hat insbesondere in der vergangenen Dekade einen massiven Aufschwung erlebt. Doch was ist eigentlich *die empirische Bildungsforschung*? Was grenzt diesen Forschungsbereich von anderen ab? Der Deutsche Bildungsrat gibt bereits im Jahr 1974 die folgende, noch immer aktuelle Definition von Bildungsforschung:

> Man kann Bildungsforschung in einem weiteren und engeren Sinne auslegen. Im engeren Sinne hat es sie als Unterrichtsforschung schon immer gegeben. Im weiteren Sinn kann sie sich auf das gesamte Bildungswesen und seine Reform im Kontext von Staat und Gesellschaft beziehen, einschließlich der außerschulischen Bildungsprozesse. Wie weit oder eng aber auch die Grenzen der Bildungsforschung gezogen werden, es sollte nur dann von Bildungsforschung gesprochen werden, wenn die zu lösende Aufgabe, die Gegenstand der Forschung ist, theoretisch oder empirisch auf Bildungsprozesse (Lehr-, Lern-, Sozialisations- und Erziehungsprozesse), deren organisatorische und ökonomische Voraussetzungen oder Reform bezogen ist. (Deutscher Bildungsrat, 1974, S. 23)

Ausgehend von dieser Definition beschreibt Gräsel (2011) drei zentrale Merkmale der empirischen Bildungsforschung: (a) Problemorientierung, d. h. wissenschaftlicher Erkenntnisgewinn mit dem Ziel, das Bildungswesen zu verbessern, (b) Interdisziplinarität und (c) Verwendung empirischer Forschungsmethoden, d. h. quantitative und qualitative Methoden.

Eine präzisere Charakterisierung der empirischen Bildungsforschung ist nicht zuletzt über deren Ziele und Gegenstandsbereiche möglich: Nach Tippelt und Schmidt (2010) besteht die Aufgabe der empirischen Bildungsforschung u. a. „ . . . darin, wissenschaftliche Informationen auszuarbeiten, die eine rationale Begründung bildungspraktischer und bildungspolitischer Entscheidungen ermöglichen“ (S. 9). Ziel ist also die Analyse und Verbesserung des Bildungswesens mittels einer *evidenzbasierten Bildungspraxis und Bildungspolitik*, bei der wissenschaftlich fundierte Ergebnisse Grundlage für bildungspolitische Entscheidungen sein sollen. Begriffe wie *Bildungsstandards* oder *Kompetenzmodelle* sind zu Schlagwörtern geworden, die in der fachlichen, politischen und öffentlichen Diskussion um Fragen der Bildung und Bildungsqualität nicht mehr wegzudenken sind. Der inhaltliche Fokus der empirischen Bildungsforschung liegt hierzulande auf den Themen Schulleistung (bspw. in Form von nationalen und internationalen Schulleistungsvergleichen) und Kompetenzentwicklung (bspw. Entwicklung individueller Fördermaßnahmen) in Schule und Hochschule (vgl. Reinders, Ditton, Gräsel & Gniewosz, 2011), was sich nicht zuletzt auch in der Ausrichtung der Bildungspolitik widerspiegelt. Im Kontext der Bildungspolitik sei beispielhaft die Gesamtstrategie zum Bildungsmonitoring (KMK, 2006) erwähnt, auf die ich in Abschnitt 2.4 eingehen

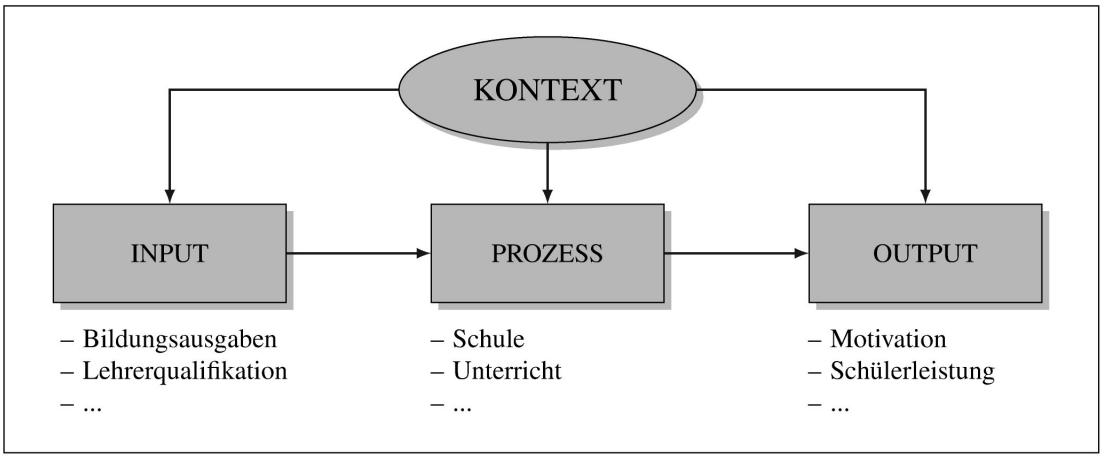

Abbildung 2.1: Basismodell der Funktionsweise von Bildungssystemen (in Anlehnung an Scheerens, 2008)

werde.

Dies war jedoch nicht immer so: Die Bedeutung der empirischen Bildungsforschung sowie einer evidenzbasierten Bildungssteuerung hat insbesondere seit dem schlechten Abschneiden deutscher Schüler in der TIMS-Studie (Baumert, Bos & Lehmann, 2000a, 2000b; Baumert et al., 1997) und dem nachfolgenden *PISA-Schock* (Deutsches PISA-Konsortium, 2001, 2002) stark zugenommen. Historisch lässt sich eine Wende von der Input- und Prozessorientierung hin zur Output-Orientierung nachzeichnen (vgl. z. B. Köller, 2010; Oelkers & Reusser, 2008). Der folgende Abschnitt skizziert diese Entwicklung überblickshaft.

2.2 Von der Input- zur Output-Orientierung

In Anlehnung an Scheerens (2008) beschreibt Köller ein „Basismodell für das Verständnis von Bildungssystemen“ (Köller, 2010, S. 529). Abbildung 2.1 zeigt ein vereinfachtes Schema dieses Modells[1]. Darin werden neben dem *Kontext* der schulischen Umwelt (bspw. bildungspolitische Maßgaben) und den *Inputs* (z. B. die jährlich investierten Bildungsausgaben) auch *Prozessvariablen* innerhalb der Schule und Klasse

[1] In der Literatur wird dieses Modell auch als *Input-Prozess-Output-Modell* bezeichnet (vgl. z. B. Müller, 2010, S. 17; Oelkers & Reusser, 2008, S. 17) Das erweiterte Modell (das sog. *integrated model of school effectiveness*), welches die einzelnen Faktoren ausdifferenziert darstellt, findet sich bei Scheerens (1990, S. 63). Ein im deutschen Sprachraum etabliertes Modell ist das *Modell zur Qualität im Bildungswesen* von Ditton (2000).

(bspw. der konkrete Unterricht in einer Klasse) als entscheidende Faktoren für den *Output* eines Bildungssystems angesehen. Letzterer umfasst die Produkte eines Bildungssystems. Dazu zählen die Leistung von Schülern bzw. – allgemein formuliert – deren Lernergebnisse, aber auch überfachliche Qualifikationen. Im Rahmen der Qualitätssicherung eines Bildungssystems werden Kriterien bzw. Standards (vgl. Abschnitt 2.3 zum Begriff der *Bildungsstandards*) festgelegt, anhand derer die Güte bzw. Qualität eines Bildungssystems beurteilt wird. Diese Kriterien können sich auf die Inputs, den Prozess oder aber die Outputs des Bildungssystems beziehen.

Nach Köller (2010) lag der Fokus der Qualitätssicherung und -entwicklung im allgemeinbildenden Schulsystem in Deutschland bis in die 1990er Jahre auf der Input- und Prozessorientierung. International vollzog sich hingegen bereits seit den 1960er Jahren ein Paradigmenwechsel von der Input- und Prozessorientierung hin zu einer Betrachtung des Outputs von Schule. Dieser Paradigmenwechsel wird auch als *empirische Wende* in den Erziehungswissenschaften bezeichnet (vgl. z. B. Köller, 2010; van Ackeren, 2002). Dies spiegelte sich einerseits in der Entwicklung und Durchführung international vergleichender, repräsentativer Schulleistungsstudien – sog. internationaler *Large Scale Assessments* – wider. Diese ermöglichen neben internationalen Vergleichen auch nationale Vergleiche. Die *International Association for the Evaluation of Educational Achievement* (IEA)[2] entwickelt seit mehr als 40 Jahren internationale Large Scale Assessments, in denen Schülerleistungen sowie Kontext- und Prozessvariablen der Bildunssysteme quantitativ erfasst und vergleichend analysiert werden. TIMSS (*Third International Mathematics and Science Study*)[3] und PIRLS/IGLU (*Progress in International Reading Literacy Study*)[4] sind zwei prominente Beispiele derartiger international vergleichender Schulleistungsuntersuchungen. Andererseits wurde diese Entwicklung in vielen Ländern durch nationale Assessments ergänzt. So wurde bspw. in den USA bereits in den 1960er Jahren das NAEP-Programm (*National Assessment of Educational Progress*)[5] initiiert – eine regelmäßig stattfindende, national repräsentative Erhebung in

[2]Die im Rahmen der vorliegenden Arbeit verwendeten Institutions-, Untersuchungs- und Fachbezeichnungen werden nach erster Aufführung in der Langform durch Abkürzungen beschrieben. Im Anhang A ist ein Verzeichnis aller Abkürzungen beigefügt.

[3]TIMSS war bis 2002 ein Akronym für *Third International Mathematics and Science Study* (Dritte Internationale Mathematik- und Naturwissenschaftsstudie). Seit dem Jahr 2003 steht das Akronym TIMSS für *Trends in International Mathematics and Science Study*.

[4]IGLU (*Internationale Grundschul-Lese-Untersuchung*) ist die nationale Bezeichnung von PIRLS.

[5]Weiterführende Informationen zum NAEP-Programm finden sich auf den folgenden Webseiten: `http://nces.ed.gov/nationsreportcard/`

ausgewählten Fächern der Klassenstufen 4, 8 und 12.

Ein anderes Bild zeigte sich in Deutschland: Zwar beteiligte sich auch die Bundesrepublik Deutschland an den großen internationalen Schulleistungsstudien der IEA (vgl. z. B. van Ackeren, 2002), jedoch – im Gegensatz zu vielen anderen Ländern – lediglich sporadisch. Außerdem stießen die Ergebnisse dieser Untersuchungen auf keine große öffentliche und politische Aufmerksamkeit. Erst in den 1990er Jahren, mit der Teilnahme an TIMSS und der Publikation der Befunde aus dieser Untersuchung, änderte sich dieses Bild auch in Deutschland. Im internationalen Vergleich lag die durchschnittliche Testleistung deutscher Schüler bezüglich der untersuchten Testbereiche Mathematik und Naturwissenschaften im unteren Mittelfeld (Baumert et al., 2000a, 2000b, 1997). Die mangelhafte Befundlage in TIMSS attestierte dem deutschen Bildungssystem lediglich Mittelmäßigkeit, was Zweifel an dessen Leistungsfähigkeit aufkommen ließ.

Durch diesen sog. *TIMSS-Schock* (van Ackeren, 2002) erhielt die systematische empirische Schulleistungsforschung einen neuen Stellenwert. Auch in Deutschland vollzog sich nun die empirische Wende in der Erziehungswissenschaft, bei der die Bedeutung der pädagogischen Psychologie und insbesondere der Psychometrie im Kontext der empirischen Bildungsforschung zunahm. Die Umorientierung von der Input- zur Output-Orientierung wurde dabei im Wesentlichen von der *Ständigen Konferenz der Kultusminister der Länder* (Kurzform: Kultusministerkonferenz, Abk.: KMK) initiiert: Als Reaktion auf das schlechte Abschneiden des deutschen Bildungssystems im Rahmen von TIMSS im Jahr 1996 stand im Rahmen der 280. Plenarsitzung der KMK im Oktober 1997 in Konstanz die Frage nach konkreten Maßnahmen zur Qualitätssicherung und Qualitätsentwicklung im Fokus. In diesem Rahmen wurden – neben der Beteiligung Deutschlands an weiteren internationalen Schulleistungsuntersuchungen – auch nationale und regionale (länderinterne) Verfahren zur Überprüfung der Lern- und Leistungsstände von Schülern als neues Instrumentarium der Qualitätssicherung diskutiert und deren Durchführung beschlossen. Im sog. *Konstanzer Beschluss* heißt es dazu:

> Die Kultusministerkonferenz sieht im Hinblick auf die Gleichwertigkeit der schulischen Ausbildung, die Vergleichbarkeit der Schulabschlüsse sowie die Durchlässigkeit des Bildungssystems innerhalb der Bundesrepublik Deutschland in der Entwicklung von Maßnahmen zur Sicherung der Qualität schulischer Bildung eine wichtige Aufgabe.

> Im Hinblick auf diese Zielsetzung und zur Qualitätssicherung in Schulen halten es die Mitglieder der Kultusministerkonferenz für erforderlich, in den Ländern Instrumente zur Evaluation zu entwickeln und zu erproben und über die gewonnenen Ergebnisse in einen breiten Erfahrungsaustausch einzutreten. (KMK, 1997, S. 1)

Die Etablierung einer evidenzbasierten Steuerung des Bildungssystems, in deren Rahmen die erreichten Schülerleistungen als ein zentrales Kriterium zur Beurteilung der Leistungsfähigkeit des Bildungssystems genutzt werden (Output-Orientierung), gewann nunmehr an Bedeutung und erreichte mit der Durchführung der PISA-Studie (*Programme for International Student Assessment*) im Jahr 2000 ihren ersten Höhepunkt. PISA ist eine international vergleichende, standardisierte Schulleistungsuntersuchung, die von der *Organisation für wirtschaftliche Zusammenarbeit und Entwicklung* (OECD) regelmäßig im Abstand von jeweils drei Jahren bei 15-jährigen Schülern durchgeführt wird (vgl. Baumert & Artelt, 2003). Teilnehmende Staaten sind neben den Mitgliedstaaten der OECD[6] eine zunehmenden Anzahl von Partnerstaaten: So nahmen im Jahr 2000 insgesamt 43 Länder an der Untersuchung teil, wohingegen es im Jahr 2009 bereits 65 Länder waren.

Das wiederum schlechte Abschneiden deutscher Schülerleistungen im Rahmen von PISA 2000 (Deutsches PISA-Konsortium, 2001, 2002) löste ein verstärktes öffentliches Interesse an Qualität und Effektivität im Bildungswesen aus. Erste Konsequenzen aus diesen Ergebnissen wurden auf der 296. Plenarsitzung der KMK im Dezember 2001 gezogen[7]. Hier werden insgesamt sieben vorrangige Handlungsfelder zur Verbesserung des deutschen Bildungssystems beschlossen. Eines dieser Handlungsfelder umfasste „Maßnahmen zur konsequenten Weiterentwicklung und Sicherung der Qualität von Unterricht und Schule auf der Grundlage von verbindlichen Standards sowie eine ergebnisorientierte Evaluation“. In der Folge nahm die Bedeutung einer evidenzbasierten Steuerung des Bildungssystems nochmals zu: Neben einer Verstetigung der Beteiligung an internationalen Schulleistungsuntersuchungen – wie bspw. an PISA oder PIRLS/IGLU – wurden die Erarbeitung der sog. *Bildungsstandards* (KMK, 2002) und weitere Evaluationsmaßnahmen auch auf Ebene der einzelnen Bundesländer (bspw. ex-

[6] Für ausführliche Informationen über die OECD sowie deren Mitglieds- und Partnerstaaten siehe: `http://www.oecd.org`

[7] Für weiterführende Informationen siehe die Pressemitteilung zur 296. Plenarsitzung der KMK: `http://www.kmk.org/presse-und-aktuelles/pm2001/296plenarsitzung.html`

terne Evaluation in Form von Schulinspektionen, Parallelarbeiten, zentrale Abschlussarbeiten oder landesweiten Vergleichsarbeiten) initiiert. Da die Bildungsstandards mittlerweile eine zentrale Rolle bei der Testentwicklung und Ergebnisrückmeldung von Vergleichsarbeiten spielen (vgl. Klieme, Avenarius et al., 2007; Klieme et al., 2010; Orth, 2007), werden diese im folgenden Abschnitt näher erläutert.

2.3 Bildungsstandards

In der internationalen Diskussion werden mindestens drei Formen von Bildungsstandards unterschieden (American Association for the Advancement of Science, 1993; National Council of Teachers of Mathematics, 2000; National Research Council, 1995): (a) *Inhaltliche Standards* (content standards) setzen Inhalte und Lernziele in einem bestimmten Fachgebiet fest. So wird bspw. im Rahmen von Lehrplänen festgelegt, welche Inhalte im Mathematikunterricht unterrichtet und von den Schülern gelernt werden sollen. (b) *Unterrichtsstandards* (opportunity-to-learn-standards) hingegen beschreiben Bedingungen des Lehrens und Lernens – bspw. in Form konsensual akzeptierter Methoden und Prinzipien gelungenen Unterrichts. Schließlich beziehen sich (c) *Leistungsstandards* (performance standards) auf die Ergebnisse – also den Output – von Lehren und Lernen, indem sie bestimmte Kompetenzen als Ziele schulischen Unterrichts definieren. Leistungsstandards lassen sich wiederum hinsichtlich der Anforderungen an die erreichten Kompetenzniveaus differenzieren in Mindest-, Regel- und Optimalstandards: *Mindeststandards* (oder auch Minimalstandards) beschreiben ein Minimum an Kompetenzen, das alle Schüler zu einem bestimmten Zeitpunkt ihrer Schullaufbahn erreicht haben sollten. *Regelstandards* hingegen definieren Kompetenzen, die im Durchschnitt von den Schülern zu einem bestimmtem Bildungsabschnitt erreicht werden sollten. Und schließlich beschreiben *Optimalstandards* (oder auch Maximalstandards) Kompetenzen, die im oberen Leistungsniveau angesiedelt sind. Sie beziehen sich also auf Kompetenzen, über die die besten Schüler einer bestimmten Jahrgangsstufe verfügen.

In der Bundesrepublik Deutschland hat man sich entschieden, die nationalen Bildungsstandards als Leistungsstandards zu definieren: Die KMK hat im Jahr 2002 beschlossen, nationale Bildungsstandards für verschiedene Jahrgangsstufen zu entwickeln. Diese legen verbindlich fest, welche Kompetenzen die Schüler einer bestimmten Klassenstufe in einem bestimmten Fach erreichen sollen (KMK, 2002). Die von der KMK

bisher vorgelegten Bildungsstandards sind als Regelstandards definiert, d. h. sie beziehen sich auf das durchschnittlich erwartete Leitungsniveau von Schülern am Ende einer bestimmten Jahrgangsstufe in einem konkreten Fachbereich (vgl. KMK, 2005). Dies unterstreicht den Paradigmenwechsel der Bildungspolitik hin zur einer Output-Orientierung. Derzeit gibt es bundesweit geltende Bildungsstandards in der Primarstufe sowie der Sekundarstufe I und II für die folgenden Fächer[8]:

- im Primarbereich (Jahrgangsstufe 4) für die Fächer Mathematik und Deutsch,
- für den Hauptschulabschluss (Jahrgangsstufe 9) für die Fächer Mathematik, Deutsch und die erste Fremdsprache (Englisch/Französisch),
- für den Mittleren Schulabschluss (Jahrgangsstufe 10) für die Fächer Deutsch, Mathematik, die erste Fremdsprache (Englisch/Französisch), Biologie, Chemie und Physik *und*
- für die Allgemeine Hochschulreife für die Fächer Deutsch, Mathematik und die fortgeführte Fremdsprache (Englisch/Französisch).

Doch was ist der Nutzen von Bildungsstandards? Bildungsstandards sollen der Bewertung und Steuerung von Bildungsprozessen dienen, indem sie verbindliche Ziele des Unterrichts in ausgewählten Fächern formulieren. Sie stellen somit „ . . . normativ gesetzte Zielgrößen dar, die in einem Bildungssystem erreicht werden sollen“ (Köller, 2011, S. 179). Um eine Bewertung und Steuerung des Bildungssystems vornehmen zu können, müssen Verfahren bzw. Messinstrumente zur Verfügung stehen, mittels derer eine valide, reliable und objektive Erfassung der erreichten Kompetenzniveaus ermöglicht wird. Parallel zur Erarbeitung der nationalen Bildungsstandards durch die KMK wurde daher im Juni 2004 das *Institut zur Qualitätsentwicklung im Bildungswesen* (IQB) gegründet. Das primäre Ziel des IQB besteht in der Operationalisierung der Bildungsstandards durch die Formulierung von Kompetenzmodellen und Generierung von Itempools, um das Erreichen der Standards überprüfen zu können (KMK, 2005). Damit wird die wissenschaftliche Begleitung bei der Implementierung der Bildungsstandards sichergestellt.

[8] Die einzelnen Dokumentationen der Bildungsstandards sind auf folgender Webseite abrufbar: `http://www.kmk.org/bildung-schule/qualitaetssicherung-in-schulen/bildungsstandards/dokumente.html`

2.4 Die Gesamtstrategie der KMK zum Bildungsmonitoring

Im Jahr 2006 schließlich legte die KMK die *Gesamtstrategie zum Bildungsmonitoring* auf Basis der Bildungsstandards vor, welche eine systematische und wissenschaftlich fundierte Evaluation von Ergebnissen des Bildungssystems verfolgt (KMK, 2006). Diese umfasst die folgenden vier Elemente, die in einem engen Zusammenspiel zu betrachten sind, jedoch jeweils verschiedene Ebenen des Bildungssystems betreffen:

(1) *Teilnahme an internationalen Schulleistungsuntersuchungen:*
Die regelmäßige Teilnahme an internationalen Schulleistungsuntersuchungen dient der Verortung der Leistungsfähigkeit des deutschen Bildungssystems im internationalen Vergleich. Die Ergebnisse dieser Untersuchungen beziehen sich also in erster Linie auf die Systemebene. Vergleichende Analysen der Ergebnisse zielen hier nicht auf die Bewertung einzelner Schulen, sondern eines gesamten Bildungssystems. Daher werden diese Untersuchungen auch als *Systemmonitoring*-Studien bezeichnet. Diese sollen „ … Systemwissen zur Verfügung [stellen], mit dem nachhaltig die Qualität eines Bildungssystems verbessert werden kann“ (Bos & Schwippert, 2002, S. 18), indem das bereitgestellte Wissen als Basis für die nationale Bildungsplanung genutzt wird. Derzeit nehmen die Bundesländer an PISA im 3-jährigen Rhythmus, PIRLS/IGLU im 5-jährigen Rhythmus und an TIMSS im 4-jährigen Rhythmus teil.

(2) *Zentrale Überprüfung des Erreichens der Bildungsstandards im Ländervergleich:*
Um neben dem internationalen Vergleich auch nationale (innerdeutsche) Vergleiche zwischen den einzelnen Bundesländern vornehmen zu können, wurden im Rahmen internationaler Schulleistungsuntersuchungen auch nationale Ergänzungsstudien durchgeführt. Hierzu zählten bpsw. PISA-E (Klieme, 2002) oder IGLU-E. Seit dem Jahr 2009 bilden die von der KMK verabschiedeten Bildungsstandards die Basis für den Vergleich der einzelnen Bundesländer. Dazu werden vom IQB länderübergreifende, standardisierte Tests entwickelt, welche die nationalen Erweiterungen wie PISA-E oder IGLU-E ablösten. Die Erhebungen für den Ländervergleich erfolgen in der Primarstufe im 5-jährigen Rhythmus und in der Sekundarstufe I im 3-jährigen Rhythmus. Der Ländervergleich – ebenso wie

die Bildungsberichtserstattung von Bund und Ländern – dient ebenfalls in erster Linie dem Systemmonitoring, denn die Ergebnisse werden auf der Ebene der Schulsysteme der einzelnen Bundesländer ausgewertet. Dabei sind keine Rückschlüsse auf die Leistungen einzelner Schulen, Klassen oder Schüler möglich oder intendiert.

(3) *Nationale Bildungsberichterstattung:*
Die gemeinsame Bildungsberichterstattung von Bund und Ländern ist ein weiterer Baustein des Bildungsmonitorings. Ziele sind u. a. eine transparente, informative Darstellung wichtiger Entwicklungen im Bildungskontext und die Rechenschaftslegung auf Ebene der einzelnen Bundesländer sowie auf Bundesebene.

(4) *Vergleichsarbeiten:*
Neben den internationalen und nationalen, länderübergreifenden Qualitätssicherungsmaßnahmen der KMK gab es bereits im Vorfeld der Gesamtstrategie verschiedene länderspezifische Maßnahmen. Zu diesen gehören neben der Implementierung der Bildungsstandards im Unterricht (bspw. durch die entsprechende Ausrichtung der Rahmenlehrpläne), die Einführung zentraler Abschlussprüfungen (bspw. des mittleren Schulabschlusses) sowie die Durchführung landesweiter Vergleichsarbeiten (vgl. Blossfeld et al., 2010). Flächendeckende Vergleichsarbeiten zur landesweiten Überprüfung der Leistungsfähigkeit einzelner Schulen, die in ausgewählten Jahrgangsstufen durchgeführt werden und mittlerweile auch an die Bildungsstandards angekoppelt werden, wurden im Jahr 2006 als weitere Säule zur Qualitätssicherung in die KMK-Gesamtstrategie zum Bildungsmonitoring aufgenommen.

Im Zentrum der vorliegenden Arbeit stehen die Vergleichsarbeiten in den Bundesländern Deutschlands, die im folgenden Abschnitt fokussiert und von anderen Evaluationsformen abgegrenzt werden.

2.5 Vergleichsarbeiten

Vergleichsarbeiten sind *standardisierte* Tests, d. h. es handelt sich um Messintrumente zur Erfassung von Schülerleistungen, die hinsichtlich psychometrischer Testgütekriterien (Objektivität, Reliabilität und Validität) entwickelt werden (Heller & Hany, 2001).

Vergleichsarbeiten sind weiterhin *standardbezogen*, da sie auf Basis der Bildungsstandards entwickelt werden bzw. sich an diesen orientieren.

Zwar wurden Vergleichsarbeiten erst im Jahr 2006 in die Gesamtstrategie der KMK zum Bildungsmonitoring aufgenommen, jedoch haben diese innerhalb der einzelnen Bundesländer eine längere, wenngleich auch jeweils unterschiedliche Tradition (vgl. Hovestadt & Kessler, 2005; van Ackeren & Bellenberg, 2004). Dies spiegelt sich u. a. in den unterschiedlichen Bezeichnungen wider: Neben dem Terminus Vergleichsarbeiten werden auch die Begriffe Lernstandserhebungen, VERA, Orientierungsarbeiten, Diagnosearbeiten, KERMIT[9] oder Kompetenztests verwendet. Im weiteren Verlauf dieser Arbeit werde ich ausschließlich den Begriff *Vergleichsarbeiten* verwenden.

Mittlerweile werden Vergleichsarbeiten bundesweit in der Primarstufe (in Klassenstufe 3) in den Fächern Mathematik und Deutsch sowie in der Sekundarstufe I (in Klassenstufe 8) in den Fächern Mathematik, Deutsch und der ersten Fremdpsrache (Englisch bzw. Französisch) durchgeführt. Zudem wurde die Testentwicklung weitestgehend vereinheitlicht und zentralisiert: Seit dem Schuljahr 2008/2009 liegt die Aufgabenentwicklung, Pilotierung und Skalierung der Items sowie die Testhefterstellung in der Verantwortung des IQB. Einzige Ausnahme hiervon bildet Baden-Württemberg, dass nicht an VERA 8 im Bundesverband teilnimmt, sondern eigene Vergleichsarbeiten für die Sekundarstufe I auf Basis der baden-württembergischen Bildungsstandards entwickelt. Die Tests beziehen sich auf zweijährige Bildungsabschnitte und werden zu Beginn der 9. Jahrgangstufe durchgeführt (vgl. Wacker & Kramer, 2012).

Trotz zunehmender Vereinheitlichung unterscheiden sich Vergleichsarbeiten nach wie vor hinsichtlich der Testdurchführung, Testkorrektur, Dateneingabe, Datenanalyse (d. h. der statistischen Auswertung der Testergebnisse) und Ergebnisrückmeldung zwischen den einzelnen Bundesländern (vgl. z. B. Fiege et al., 2011), da diese Aufgaben in der Verantwortung der einzelnen Länder liegen. Die damit einhergehende Heterogenität der Vorgehensweisen zeigt sich zusätzlich in den Zielen und Funktionen, die Vergleichsarbeiten zugeschrieben werden.

[9] KERMIT ist ein Akronym für **K**ompetenzen **ermit**teln. KERMIT ist ein System zur Erfassung der Kompetenzentwicklung der Schüler, das seit dem Schuljahr 2012/2013 in Hamburg etabliert wird. Dieses umfasst neben den Vergleichsarbeiten in Klassenstufe 3 und 8 zusätzlich die in Hamburg etablierten Lernausgangslagenerhebungen in den Klassenstufen 2, 5, 7 und 9.

2.5.1 Ziele und Funktionen von Vergleichsarbeiten

In dem Beschluss der KMK-Gesamtstrategie aus dem Jahr 2006 heißt es: „Vergleichsarbeiten dienen insbesondere der flächendeckenden, jahrgangsbasierten Evaluation der einzelnen Schule und Klasse als Standortbestimmung vor dem Hintergrund der länderübergreifenden Bildungsstandards“ (S. 10). Landesweite Vergleichsarbeiten dienen somit der Überprüfung des Erreichens der Bildungsstandards auf Einzelschulebene: Da die Tests auf Basis der Bildungsstandards entwickelt werden, kann die einzelne Schule bzw. Klasse die Ergebnisse nutzen, um die Stärken und Schwächen hinsichtlich des Erreichens der Bildungsstandards zu diagnostizieren. Vergleichsarbeiten sollen auf diese Weise quantifizieren, in welchem Maße die Schüler über die in den Bildungsstandards festgelegten Kompetenzen verfügen. Dies soll zur Qualitätsentwicklung innerhalb der einzelnen Schulen beitragen, indem „Defizitbereiche und Handlungsbedarfe“ (vgl. van Ackeren & Klemm, 2009, S. 164) aufgezeigt werden. Neben dieser Standortbestimmung geht es gleichfalls – explizit oder nur implizit formuliert – um die Evaluation der Wirksamkeit schulischer Arbeit, also um die Frage: Wie gut bzw. wie effektiv ist der Unterricht in einer Klasse?

Weitere Funktionen von Vergleichsarbeiten und die Ebenen des Bildungssystems

Seit dem Jahr 2004 gibt es das *Netzwerk zur empiriegestützten Schulentwicklung* (EMSE). EMSE ist ein kooperatives Netzwerk bestehend aus Mitgliedern der Kultusministerien, Qualitätsagenturen und Landesinstituten aus allen 16 Bundesländern. In den regelmäßig stattfindenden Fachtagungen werden u. a. aktuelle Forschungsergebnisse zum Thema empirische Schulentwicklung und Bildungsplanung rezipiert und hinsichtlich praxisrelevanter Konsequenzen diskutiert. Einen wichtigen Stellenwert nimmt dabei insbesondere der Erfahrungsaustausch zwischen den Ländern im Hinblick auf Maßnahmen empirisch orientierter Schul- und Unterrichtsentwicklung ein. So beschreibt das EMSE-Netzwerk in seinem ersten Positionspapier aus dem Jahr 2006 die verschiedenen Funktionen von zentralen standardisierten Lernstandserhebungen – also Vergleichsarbeiten – auf den unterschiedlichen Ebenen des Bildungssystems. Tabelle 2.1 fasst diese Aufgaben zusammen. Vergleichsarbeiten werden somit Funktionen auf der Systemebene (z. B. Ressourcenallokation), Schul- und Klassenebene (z. B. Informationen über die Wirksamkeit schulischer Arbeit) *und* der Ebene individueller Schüler (bspw. Lern-

Tabelle 2.1: Funktionen von Vergleichsarbeiten differenziert nach den Ebenen des Bildungssystems

Ebene	**Funktion**
Systemebene	– Informationen über Leistungen des Bildungssystems (differenzierbar nach u. a. Unterrichtsfach, Schulform, Klassenstufe) – Gewinnung von Planungs- und Steuerungswissen (bspw. Implikationen für die Schulsystementwicklung) – Identifizierung von Unterstützungs- und Entwicklungsbedarf (bspw. regionalisierter Support, Ressourcenallokation)
Schul- & Klassenebene	– Informationen über erreichte Lernstände sowie über die Wirksamkeit der schulischen Arbeit (Unterrichts- und Schuleffekte) – Qualitätssicherung & Unterrichtsentwicklung – Rechenschaftslegung & Transparenz (z. B. Information für Eltern)
Schülerebene	– Informationen über erreichte Lernstände der Schüler (Individuelle Standortbestimmung) – Lernbedarfsdiagnosen – Rückmeldungen und Dialog mit Schülern und Eltern (Feedback-Kultur) – Zertifizierung[a]

Anmerkungen. Adaptiert aus dem EMSE-Positionspapier (vgl. EMSE, 2006, S. 1–2).
[a] Bis zum Schuljahr 2011/2012 wurden in einigen Bundesländern Vergleichsarbeiten als Klassenarbeit gewertet und benotet.

bedarfsdiagnosen) zugeschrieben.

Auch Lorenz (2005) arbeitete verschiedene Ziele bzw. Funktionen aus, die sich mittels der Ergebnisse aus Vergleichsarbeiten erreichen lassen: So sollen Vergleichsarbeiten in der Primarstufe der Verbesserung von Chancengleichheit durch Objektivierung (Hilfe zur Selbstevaluation), als Orientierungshilfe für die Schullaufbahnberatung und der Stärkung der diagnostischen Kompetenzen der Lehrkräfte dienen. Weitere Funktionen liegen nach Lorenz (2005) in der Überprüfung des Erreichens der Bildungsstandards.

Und nicht zuletzt lässt sich in den einzelnen Bundesländern – u. a. auf den entsprechenden Webseiten bspw. der Landesbildungsserver – ein Kaleidoskop von intendierten Zielen und Funktionen finden, von denen hier nur einige exemplarisch herausgegriffen

werden sollen: In Bayern wird der pädagogische Nutzen von Vergleichsarbeiten neben der Lernstandsdiagnostik in der individuellen Förderung gesehen. Ergebnisse aus Vergleichsarbeiten sollen weiterhin einen potentiellen Austausch und die Kooperation mit den Eltern unterstützen. Schließlich sollen Vergleichsarbeiten zur Schul- und Unterrichtsentwicklung durch die Evaluation der „. . . Wirksamkeit der eigenen [schulischen] Arbeit“ (Hausknecht & Eyrainer, 2010, S. 11) beitragen. In Nordrhein-Westfalen werden zusammenfassend folgende Ziele formuliert[10]: „Feststellung des Lern- und Förderbedarfs in den überprüften fachlichen Bereichen; Weiterentwicklung des Unterrichts und der schulischen Arbeit; Standardüberprüfung und Qualitätssicherung; Unterstützung der Umsetzung der Kernlehrpläne und nationalen Bildungsstandards; Stärkung der diagnostischen Kompetenz von Lehrkräften *und* Bereitstellung von ergänzenden Informationen für die schulübergreifende Qualitätssicherung“. In Thüringen sollen die „ . . . Tests [. . .] eine schulische Selbstevaluation bezüglich der Wirkungsqualität“ ermöglichen[11].

Vereinbarung der KMK zur Weiterentwicklung von Vergleichsarbeiten

Zwar lassen sich die Funktionen von Vergleichsarbeiten allen drei Ebenen – Systemebene, Schul- und Klassenebene, Schülerebene – zuordnen (vgl. Tabelle 2.1), jedoch liegt der Fokus auf der Schul- und Klassenebene (z. B. Nachtigall, Kröhne, Enders & Steyer, 2008; Nachtigall, Storbeck & Landmann, 2009): Nach Isaac und Hosenfeld (2008) sollen Vergleichsarbeiten in erster Linie „ . . . fachliche, fachdidaktische und pädagogisch-psychologische Impulse für Schul- und Unterrichtsentwicklung liefern“ (S. 144). Dieser Fokus ist auch von der KMK im Jahr 2012 in der *Vereinbarung zur Weiterentwicklung von VERA* bekräftigt worden (KMK, 2012). Dort heißt es in der Zielbestimmung von Vergleichsarbeiten in den Ländern: „Die zentrale Funktion von VERA als einem von vier Elementen der Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring liegt in der Unterrichts- und Schulentwicklung jeder einzelnen Schule“ (KMK, 2012, S. 2). Die Zielbestimmung wurde weiter präzisiert, indem explizit formuliert wurde, welche Aufgaben Vergleichsarbeiten *nicht* erfüllen sollen oder können: Ergebnisse aus Vergleichsarbeiten sollen fortan nicht von den Bildungs-

[10] Informationsseiten des Landes Nordrhein-Westfalen zum Lernstand 8 (Stand: 01.03.2013): http://www.standardsicherung.schulministerium.nrw.de/lernstand8/aktuelles/

[11] Informationsseiten des Landes Thüringen zu den Kompetenztests (Stand: 01.03.2013): http://www.thueringen.de/th2/tmbwk/bildung/information/kompetenztest/

Tabelle 2.2: Vergleich verschiedener Evaluationsformen im Schulkontext (in Anlehnung an van Ackeren und Klemm, 2009)

Evaluationsform	Funktionen			
	Systemmonitoring	Rechenschaftslegung	Zertifizierung/ Selektion	Diagnose von Stärken & Schwächen
Parallelarbeiten		×	×	×
Zentrale Abschlussprüfungen		×	×	
Vergleichsarbeiten	×	×	(×)[a]	×
Nationale Schulleistungsstudien[b]	×	×		
Internationale Schulleistungsstudien[c]	×	×		

Anmerkungen. [a] Bis zum Schuljahr 2011/2012 wurden Vergleichsarbeiten in einigen Bundesländern als Klassenarbeit gewertet und benotet.
[b] Hierzu zählt bspw. der KMK-Ländervergleich, der seit dem Schuljahr 2008/2009 in der Primarstufe alle fünf Jahre und in der Sekundarstufe I alle drei Jahre durchgeführt wird.
[c] Hierzu zählen u. a. PISA, PIRLS/IGLU und TIMSS.

verwaltungen der Länder veröffentlicht werden. Weiterhin sollen diese weder benotet werden noch für Übergangsempfehlungen eingesetzt werden.

2.5.2 Vergleichsarbeiten als spezielle Form der Evaluation

Vergleichsarbeiten stellen eine spezielle Evaluationsform im Bildungskontext dar. Evaluation ist ein Prozess, „ … in dessen Verlauf eine *Bestandsaufnahme*, eine *Analyse* und eine *Bewertung*“ (van Ackeren & Klemm, 2009, S. 159) eines Untersuchungsgegenstandes – z. B. der Arbeit einer Schule, der Qualität des Unterrichts einer Klasse, der Kompetenzen eines Schülers etc. – erfolgt. Ausgehend von dieser sehr allgemeinen Definition lassen sich im Bildungskontext eine Reihe unterschiedlicher Evaluationsformen unterscheiden. So grenzen van Ackeren und Klemm (2009) Parallelarbeiten, zentrale Abschlussprüfungen und Schulleistungsuntersuchungen von Vergleichsarbeiten hinsichtlich der jeweiligen Funktionen ab (vgl. Tabelle 2.2). Einen differenzierteren Überblick der verschiedenen Evaluationsformen bzw. Formen von Schulleistungsuntersuchungen im Kontext der deutschen Bildungslandschaft findet sich in Anhang B. Den aufgezählten Evaluationsformen ist gemeinsam, dass sie Ausgangspunkt von Steue-

rungs- bzw. Veränderungsprozessen im Bildungskontext sein sollen. Während Parallelarbeiten dabei insbesondere Veränderungen auf der Klassen- und ggf. auch auf Schulebene intendieren, liegt der Fokus nationaler und internationaler Schulleistungsuntersuchungen auf der Systemebene. Vergleichsarbeiten nehmen diesbezüglich eine Sonderstellung ein: Zwar soll der Fokus auf der Klassen- und Schulebene liegen, jedoch ermöglicht ihre Konzeption als standardisierte und standardbezogene Tests zusätzlich Vergleiche und Bewertungen auf der Systemebene. Die Vergleichsfunktion wird nicht zuletzt auch durch den Terminus *Vergleichsarbeiten* nahelegt. Diese Funktion der besseren Vergleichbarkeit derartiger Testergebnisse wird bspw. auch vom *Staatsinstitut für Schulqualität und Bildungsforschung* (ISB) in einer Handreichung für die Umsetzung von Vergleichsarbeiten an bayrischen Schulen explizit benannt. Dort heißt es: „Der Blick von außen auf die eigene Klasse ist hilfreich und ermöglicht empirisch gesicherte Vergleiche, die einem ansonsten nicht zur Verfügung stehen: den Vergleich mit anderen Klassen und Schulen“ (Hausknecht & Eyrainer, 2010, S. 10).

Ein gemeinsames Ziel von landesweiten Vergleichsarbeiten ist die Evaluation von Unterrichtseffekten auf der Ebene einzelner Schulklassen, also des Ergebnisses von Lehre und Unterricht: Basierend auf den Testergebnissen der Schüler in den Vergleichsarbeiten sollen Maßnahmen zur Unterrichts- und Qualitätsentwicklung erarbeitet werden können, die der Lehrer einer Klasse nutzen kann. Vergleichsarbeiten können somit als spezielle Evaluationsform betrachtet werden, die durch die Quantifizierung von Unterrichtseffekten Ausgangspunkt für Veränderungen im pädagogischen Handeln von Lehrern sein soll. Damit eine Evaluation Ausgangspunkt von Veränderungen sein kann, müssen verschiedene Bedingungen erfüllt sein, die jeweils unterschiedliche Komponenten des Evaluationsprozesses betreffen. Abbildung 2.2 zeigt ein vereinfachtes Schema des Evaluationsprozesses im Kontext von Vergleichsarbeiten, das drei wesentliche Elemente darstellt (vgl. Fiege et al., 2011):

(1) *Die Messung der Schülerleistung:*
Die erste Komponente umfasst die empirische Erfassung von Schülerleistungen bzw. -kompetenzen und von weiteren Merkmalen des Lernumfeldes[12] wie bspw. Eigenschaften der Schüler. Bei diesem erste Schritt im Evaluationsprozess stehen psychometrische Testgütekriterien im Vordergrund: Die zuverlässige, valide und objektive Messung von Schülerleistungen in verschiedenen Anforderungsberei-

[12] Diese weiteren Merkmale des Lernumfeldes werde ich nachfolgend als *Kovariaten* bezeichnen.

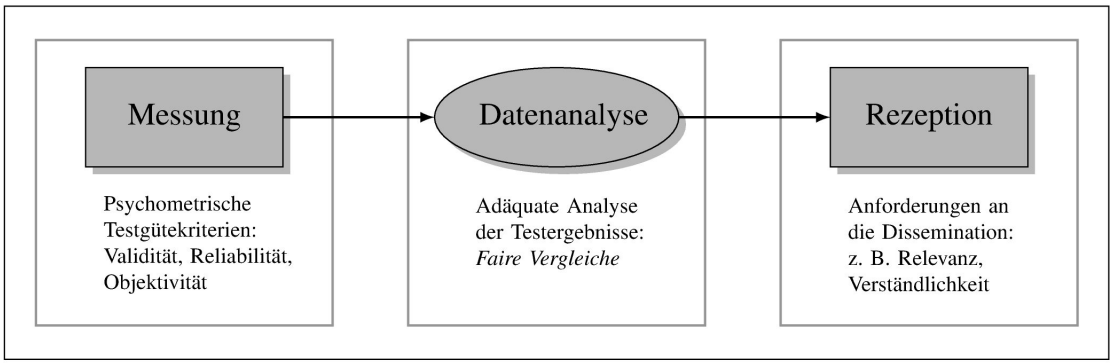

Abbildung 2.2: Essentielle Komponenten des Evaluationsprozesses im Kontext von Vergleichsarbeiten (in Anlehnung an Fiege et al., 2011)

chen sowie von Kontextbedingungen des Lernens ist eine wesentliche Basis für die Vergleichbarkeit von Bildungsergebnissen. Ein großer Anteil aktueller Forschungsbemühungen betrifft die Kompetenzmessung (z. B. Hartig & Frey, 2012; Hartig & Klieme, 2006; Hartig, Klieme & Leutner, 2008; Klieme & Hartig, 2008; Klieme & Leutner, 2006; Leucht, Harsch, Pant & Köller, 2012; Weinert, 2002).

(2) *Die Analyse der anfallenden Daten:*
Resultat der Messung sind im Rahmen von Vergleichsarbeiten quantitative Daten. Ein zweiter wichtiger Schritt, um die Effekte des Unterrichts zu bestimmen, ist die Analyse der Daten. Die Wahl geeigneter statistischer Analyseverfahren ist ausschlaggebend für die Interpretierbarkeit der Ergebnisse als Unterrichtseffekte. Ein *fairer Vergleich* (vgl. Abschnitt 2.5.4) ist dabei die notwendige Voraussetzung zur validen Einschätzung von Unterrichtseffekten, d. h. der Wirkung von Unterricht.

(3) *Die Rezeption der Ergebnisse:*
Die dritte wichtige Komponente ist die Rezeption der Ergebnisse seitens der Akteure im Bildungssystem, d. h. durch Lehrkräfte, Schulleiter, Schüler und teilweise auch Eltern. Die Ergebnisrezeption und -interpretation erfolgt im Kontext von Ergebnisrückmeldungen durch ein im Allgemeinen methodisch wenig geschultes Publikum, wobei die primäre Zielgruppe im Rahmen von Vergleichsarbeiten die Lehrkräfte der einzelnen Klassen sind (vgl. Fiege, 2007). Werden klassenbezogene Rückmeldungen den beteiligten Lehrpersonen zur Verfügung gestellt (*Dissemination*), so kann nach (Rolff, 2002) nicht unweigerlich davon ausgegan-

gen werden, dass hieraus unmittelbar Maßnahmen zur Unterrichtsentwicklung folgen. Erst wenn die Möglichkeiten und Grenzen der Interpretation dieser Ergebnisse verstanden werden, können Veränderungen der pädagogischen Arbeit in den Schulen erwartet werden. Ergebnisse aus der Rezeptionsforschung bei Vergleichsarbeiten, die erst in überschaubarer Anzahl vorliegen (z. B. Helmke & Hosenfeld, 2005; Kuper & Schneewind, 2006; Maier, 2009; Meyer, 1997; Nachtigall et al., 2009; Wacker & Kramer, 2012), können somit wichtige Impulse in der Weiterentwicklung dieses Evaluationsintrumentariums liefern.

Zusammenfassend sind alle drei dargestellten Komponenten gleichermaßen wichtig. Im Zentrum der vorliegenden Arbeit steht die mittlere Komponente (Datenanalyse), denn die adäquate Analyse der quantitativen Daten bildet eine zentrale Gelenkstelle bei der Evaluation von Unterrichtseffekten.

2.5.3 Bezugsnormen bei der Leistungsbewertung: Womit vergleichen Vergleichsarbeiten?

Die reine Erhebung von Leistungsdaten – d. h. die Messung – mittels standardisierter Testverfahren ist nicht ausreichend zur Evaluation von Unterrichtseffekten. Zur Beurteilung der aus den Vergleichsarbeiten resultierenden Testleistungen ist ein Vergleich dieser mit einem Kriterium, einer Referenz bzw. einem Standard erforderlich. Hierfür lassen sich drei Arten von Vergleichsstandards, den sog. *Bezugsnormen* differenzieren, die jeweils unterschiedliche Informationen bereitstellen: (a) die individuelle, (b) die kriteriale und (c) die soziale Bezugsnorm (Rheinberg, 2001). Alle drei lassen sich im Kontext der Ergebnisrückmeldung aus Vergleichsarbeiten finden (Helmke & Hosenfeld, 2004; Helmke, Hosenfeld & Schrader, 2004).

Individuelle Bezugsnorm

Bei der individuellen Bezugsnorm – synonym auch *ipsative* Bezugsnorm – werden verlaufsorientierte bzw. entwicklungsbezogene Vergleiche angestellt, d. h. es wird die Veränderung der Leistung einer Klasse[13] bewertet. Zu diesem Zweck wird die Leistung zu

[13] Die Analyseeinheit können Schüler, Klassen, Schulen, Länder etc. sein. Dies ist unabhängig von der Wahl der jeweiligen Bezugsnorm. Im Folgenden verwende ich beispielhaft die Klassenebene, da diese im Fokus der vorliegenden Arbeit steht.

einem früheren Zeitpunkt als Referenz genutzt. Es geht somit um die intraindividuelle Leistungsveränderung über die Zeit.

Kriteriale Bezugsnorm

Hierbei dient ein inhaltliches Kriterium zur Beurteilung der Leistung. Kriteriale Vergleiche beziehen sich auf die Frage: Wie ist die Leistung einer Klasse im Vergleich zu einem zuvor festgelegten inhaltlichen Kriterium zu beurteilen? So wurden z. B. im Rahmen der PISA-Untersuchung (Deutsches PISA-Konsortium, 2001) voneinander abgrenzbare Kompetenzstufen inhaltlich definiert. Die Einordnung einer Schülerleistung in eine der Kompetenzstufen beschreibt das Kompetenzniveau dieses Schülers mittels charakteristischer Fähigkeiten, die mit dieser Kompetenzstufe assoziiert sind. Diese Vergleichsperspektive hat mit der Einführung der nationalen Bildungsstandards (KMK, 2005) an Bedeutung gewonnen, vor deren Hintergrund gleichfalls Kompetenzstufen modelliert werden. Dabei wird die kontinuierliche Kompetenzskala, die mittels der Vergleichsarbeiten gemessen wird, kategorisiert: Mittels des sog. *Standard-Setting-Verfahrens* werden mehrere Schwellenwerte (Cut-Scores) bestimmt, „ . . . durch die benachbarte Kategorien auf einer kontinuierlichen Testwertskala abgegrenzt werden“ (Pant, Tiffin-Richards & Köller, 2010, S. 175). Die dabei resultierenden Abschnitte werden als Kompetenzstufen bezeichnet. Die klassenspezifischen Rückmeldungen der Ergebnisse aus den Vergleichsarbeiten enthalten seither zumeist auch die Einordnung der Schülerleistungen hinsichtlich der Kompetenzstufen.

Die Kompetenzstufenverteilung einer Klasse – d. h. die relativen Häufigkeiten der erreichten Kompetenzstufen aller Schüler einer Klasse – ermöglicht jedoch zunächst keine Aussagen über Unterrichtseffekte. Um darüber Aussagen treffen zu können, benötigt man längsschnittliche Informationen, die einen Vergleich der aktuellen mit der Kompetenzstufenverteilung dieser Klasse zu einem früheren Zeitpunkt ermöglichen. Alternativ dazu kann man die Kompetenzstufenverteilung dieser Klasse mit der einer anderen (vergleichbaren) Klasse vergleichen. Letzteres ist ein sozialer Vergleich mit kriterialen Normen.

Der Nachteil von kriterialen Vergleichen liegt somit darin, dass diese für sich genommen keine Informationen über Unterrichtseffekte enthalten. Dies ist nur dann gegeben, wenn man eine weitere Vergleichsdimension berücksichtigt. Ein Beispiel dafür ist die Kombination von kriterialer und sozialer Bezugsnorm.

Soziale Bezugsnorm

Im Gegensatz zu kriterialen Vergleichen erfolgt die Beurteilung einer Schülerleistung im Rahmen sozialer Vergleiche auf Basis der Leistungsverteilung aller Schüler, deren Leistung erhoben wurde. Damit zielen soziale Vergleiche auf die Beantwortung der Frage: Wie ist die Leistung einer Klasse im Vergleich zur Leistung anderer Klassen, die von anderen Lehrkräften unterrichtet wurden, zu beurteilen?

Die soziale Bezugsnorm ist besonders geeignet, um Aussagen über den Effekt des Unterrichts in einer Klasse im Vergleich zum Unterricht in anderen Klassen zu treffen, denn ein Effekt ist stets der Unterschied zwischen zwei Bedingungen (vgl. Holland, 1986). Bei den im Kontext von Vergleichsarbeiten betrachteten Bedingungen handelt es sich um die verschiedenen Unterrichtsbedingungen. Mit anderen Worten formuliert: Indem ein Lehrer die durchschnittliche Leistung der eigenen Klasse mit dem Leistungs-Output anderer vergleichbarer Klassen, die nicht von ihm unterrichtet wurden, vergleicht, kann er beurteilen, wie viel besser bzw. schlechter seine Schülerschaft in Folge seines Unterrichts ist. Die Betonung liegt hierbei auf dem Ausdruck *vergleichbare Klassen*, denn die notwendige Voraussetzung für die Interpretation als Unterrichtseffekte sind sog. *faire Vergleiche*, welche systematische, nicht auf den Unterricht attribuierbare Unterschiede zwischen Klassen berücksichtigen.

2.5.4 Faire Vergleiche in Vergleichsarbeiten

Auch in Vergleichsarbeiten wird die soziale Bezugsnorm als Referenz zur Beurteilung der Wirksamkeit des Unterrichts verwendet: So enthalten die klassenspezifischen Ergebnisrückmeldungen neben dem Testergebnis häufig den Landesdurchschnitt. Zudem können die Ergebnisse zusätzlich zwischen den verschiedenen Klassen einer Schule verglichen werden.

Die dabei resultierenden Unterschiede sind jedoch nicht allein auf die Effektivität des Unterrichts zurückzuführen, sondern in starkem Maße auch von weiteren Faktoren beeinflusst (z. B. Isaac & Hosenfeld, 2008; Nachtigall & Kröhne, 2006; Nachtigall et al., 2008). Derartige außerschulische Einflussfaktoren auf das Lernen sind u. a. das Vorwissen, der sozioökonomische Status, das Geschlecht, die Muttersprache oder auch die soziale Komposition der Klasse eines Schülers. Diese Variablen, die ich nachfolgend als *Kovariaten* bezeichne, können folglich sowohl individuelle Schülermerkmale reprä-

sentieren, als auch Kontextvariablen auf Klassen- oder/und Schulebene. Gemeinsames Merkmal von Kovariaten ist, dass sie dem Unterrichtsprozess bzw. dem pädagogischen Handeln eines Lehrers zeitlich vorgeordnet sind. Somit sind Kovariaten von dem Lehrer bzw. der Schule nicht beeinflussbar, haben jedoch ihrerseits Effekte auf die Schülerleistung. Ein einfacher Mittelwertsvergleich – z. B. ein Vergleich mit dem Landesdurchschnitt – berücksichtigt solche Kontextvariablen des Lernens nicht und verfehlt damit das Ziel, die Effektivität des Unterrichts zu quantifizieren. Es besteht daher ein allgemeiner Konsens, dass für *faire Vergleiche* statistische Adjustierungsverfahren verwendet werden müssen, um diesen Unterschieden Rechnung zu tragen (z. B. Baumert, Stanat & Watermann, 2006; OECD, 2008; Watermann & Stanat, 2004; Wegscheider, 2004).

Statistische Adjustierungsverfahren zielen folglich auf die Beantwortung der Frage: Welches Testergebnis hätte eine Klasse unter sonst gleichen Ausgangsbedingungen erzielt, wenn eine andere Lehrperson den Unterricht in dieser Klasse gestaltet hätte (*Ceteris-paribus-Klausel*; vgl. Mill, 1865). Die Differenz dieses adjustierten Wertes zum tatsächlichen Testergebnis einer Klasse kann dann ursächlich auf den Effekt des Unterrichts attribuiert werden. Ein *fairer* Vergleich ist demnach nur dann fair, wenn dieser kausal interpretierbar ist, d. h. wenn er den Effekt des Unterrichts in einer Klasse auf die Schülerleistung quantifiziert. Eine solche kausale Attribution ist jedoch nur unter bestimmten Bedingungen möglich. Um diese Bedingungen explizieren zu können, bedarf es eines theoretischen Rahmens (vgl. Fiege, 2007). Im Rahmen dieser Arbeit verwende ich dazu eine allgemeine stochastische Theorie kausaler Effekte (vgl. Steyer et al., 2011).

2.6 Zusammenfassung

In Deutschland lässt sich – insbesondere im Verlauf der vergangenen Dekade – eine Entwicklung hin zu einer evidenzbasierten Bildungssteuerung nachzeichnen: Bildungspolitische Entscheidungen sollen zunehmend auf Basis empirischer Befunde getroffen bzw. von diesen mitbestimmt werden. Des Weiteren sollen Maßnahmen zur Qualitätssicherung und -entwicklung auf empirischen Befunden beruhen. Dieser Trend, der mit einer empirischen Wende in den Erziehungswissenschaften einhergeht, kann in anderen Ländern bereits sehr viel früher nachgewiesen werden. Die historischen Hintergründe

dieser Entwicklung wurden im vorliegenden Kapitel skizziert.

Vergleichsarbeiten sind mittlerweile ein etabliertes Werkzeug zur empirischen Qualitätsentwicklung und -sicherung im Bildungswesen. Diese sind seit dem Jahr 2006 Bestandteil der Gesamtstrategie zum Bildungsmonitoring der Kultusministerkonferenz, die von insgesamt vier Säulen getragen wird. Vergleichsarbeiten sind mit vielen, teilweise äußerst heterogenen Zielstellungen verknüpft, die keinesfalls gleichzeitig mit ein und demselben Evaluationsinstrument erreicht werden können. Nicht zuletzt aus diesem Grund hat die KMK im Jahr 2012 im Rahmen der Weiterentwicklung der Vergleichsarbeiten vereinbart, dass die zentrale Funktion in der Unterrichts- und Schulentwicklung der einzelnen Schulen liegen soll.

Vergleichsarbeiten sind standardbezogene und standardisierte Messinstrumente. Die Ergebnisse dieser standardisierten Testverfahren sollen Ausgangspunkt von Unterrichtsentwicklung sein. Zu diesem Zweck werden die Testergebnisse einer Klasse u. a. mit den Ergebnissen anderer Klassen oder Schulen verglichen (soziale Bezugsnorm). In der KMK-Vereinbarung zur Weiterentwicklung der Vergleichsarbeiten aus dem Jahr 2012 heißt es dazu: „Die Länder streben eine Ergebnisrückmeldung mittels geeigneter Referenzgruppen (sogenannter ‚Fairer Vergleich') auf Schulebene unter Wahrung landesspezifischer Gesetze und Richtlinien (Datenschutz) an“ (KMK, 2012, S. 4). Um faire Vergleiche und damit Aussagen über die Wirksamkeit schulischer Arbeit zu ermöglichen, müssen hier außerschulische Einflussfaktoren auf das Lernen (Kovariaten) in der statistischen Auswertung der Testergebnisse berücksichtigt werden.

Um den Begriff der *Fairness* bei Leistungsvergleichen klarer abgrenzen zu können und darauf aufbauend Adjustierungsverfahren hinsichtlich der Fairness ihrer Ergebnisse bewerten zu können, bedarf es eines theoretisches Fundaments. In der vorliegenden Arbeit wird die Theorie kausaler Effekte in der Tradition von Neyman und Rubin (Neyman, 1923/1990; Rubin, 1974, 1977, 1978; Steyer et al., 2011) als theoretischer Ansatzpunkt dienen, auf dessen Basis statistische Adjustierungsverfahren bei Vergleichsarbeiten hinsichtlich der Fairness ihrer Ergebnisse beurteilt werden sollen. Zu diesem Zweck werden im nachfolgenden Kapitel die zentralen Konzepte und Annahmen einer allgemeinen stochastischen Theorie kausaler Effekte (Steyer et al., 2011) dargestellt.

Es gibt nichts Praktischeres als eine gute Theorie.

KURT LEWIN (1890 – 1947)

3 Kausale Effekte: Faire Vergleiche und die Theorie kausaler Effekte

Ebenso wie in anderen empirischen Wissenschaftsdomänen kommt der Analyse kausaler Effekte auch innerhalb der empirischen Bildungsforschung eine herausragende Rolle zu. Kausale Zusammenhänge, welche Ursache und Wirkung in Beziehung setzen, sind häufig – wenn auch oftmals nicht explizit – Ausgangspunkt von Hypothesengenerierung und -prüfung, theoretischer Modellbildung und schließlich auch der Ableitung konkreter Handlungskonsequenzen im Anwendungskontext. Bildungspolitische Entscheidungen, die eine Veränderung des Bildungssystems betreffen und mit erheblichen finanziellen sowie zeitlichen Kosten verbunden sind, lassen sich nur dann rechtfertigen, wenn sie die intendierte Wirkung zeigen. Die Entscheidung über die Abschaffung des gegliederten Schulsystems oder die Einführung eines speziellen Förderprogramms für hochbegabte Schüler sind nur zwei Beispiele derartiger Veränderungsmaßnahmen. Oft sollen empirische Untersuchungen im Rahmen von Evaluationen die Wirkung solcher Maßnahmen überprüfen. Es besteht jedoch ein allgemeiner Konsens darüber, dass nicht jede empirische Untersuchung kausale Schlussfolgerungen ermöglicht. Wohlbekannt ist in diesem Zusammenhang der Ausdruck „correlation is not causation“ (z. B. Barnard, 1982). Um Bedingungen und Grenzen von kausalen Inferenzen explizit zu machen, ist daher ein theoretisches Fundament unverzichtbar. Vor diesem Hintergrund soll im Folgenden eine *allgemeine stochastische Theorie kausaler Effekte* (Steyer et al., 2011) vorgestellt werden.

In der allgemeinen stochastischen Theorie kausaler Effekte werden die zentralen Begriffe, die in empirischen Anwendungen und Fragestellungen von Interesse sein können, definiert. Die der Theorie zugrunde liegenden Ideen sind jedoch keineswegs neu, sondern wurden bereits im Verlauf des 20. Jahrhunderts in der statistischen Literatur konzipiert und fortentwickelt. Diese Entwicklung geht u. a. zurück auf Sir Ronald A. Fisher (1946), der die Randomisierung als Versuchsplanungstechnik einführte, oder

auch Jerzy Neyman (1923/1990), auf den das Konzept des *individuellen kausalen Effekts* im Kontext von Agrar- bzw. Landwirtschaftsforschung zurückzuführen ist. Neymans Ideen und Konzepte wurden später von Donald B. Rubin (z. B. Rubin, 1974, 1977, 1978) aufgegriffen und weiterentwickelt. Die stochastische Theorie kausaler Effekte nach Steyer et al. (2011) stellt ebenfalls eine Weiterentwicklung bzw. Verallgemeinerung dieser früheren Theorien kausaler Effekte dar.

Im folgenden Kapitel soll ein Überblick über die zentralen Konzepte und Annahmen der allgemeinen stochastischen Theorie kausaler Effekte in der Neyman-Rubin-Tradition (Steyer et al., 2011) gegeben werden[1]. Dabei soll gleichfalls der Bezug der theoretischen Begriffe zum Anwendungskontext der Schulleistungsuntersuchungen bzw. insbesondere zu Vergleichsarbeiten, auf denen der Fokus der vorliegenden Arbeit liegt, hergestellt werden. Die folgenden Ausführungen basieren auf Steyer et al. (2011) sowie Fiege (2007).

3.1 Gegenstandsbestimmung

Bevor die theoretischen Konzepte eingeführt werden, ist es zunächst notwendig, eine Abgrenzung des Kausalitätsbegriffes – des Gegenstandes der Theorie kausaler Effekte – vorzunehmen. Dies ist nicht zuletzt deshalb wichtig, da dieser im wissenschaftlichen Diskurs mit unterschiedlichen Bedeutungen belegt sein kann. So führt bspw. Holland (1986) an: „One difficulty that arises in talking about causation is the variety of questions that are subsumed under the heading“ (S. 945). Was nun aber ist der zentrale Inhaltsbereich der Theorie?

Steyer et al. (2011) benennen verschiedene Aspekte, welche die empirische Kausalforschung charakterisieren. Hierzu zählt einerseits die *statistische Inferenz*, d. h. das Schließen von Stichprobenkennwerten auf die entsprechenden Populationsparameter[2]. Dies betrifft bspw. die Inferenz von der (u. U. messfehlerbehafteten) Mittelwertsdifferenz $\hat{Y}_x - \hat{Y}_{x'}$ bezüglich einer Outcome-Variablen Y zwischen zwei Gruppen x und

[1] Die nachfolgenden Ausführungen basieren auf dem Stand der Theorie im Jahr 2011. Marginale Abweichungen – insbesondere bezüglich der formalen Darstellung der theoretischen Größen – zwischen der Darstellung in diesem Kapitel und der aktuellen Version der allgemeinen stochastischen Theorie kausaler Effekte sind daher nicht ausgeschlossen und konnten aufgrund der zeitlichen Limitation beim Verfassen dieser Qualifikationsarbeit nicht gänzlich vermieden werden. Die aktuelle Version der Theorie findet sich in Steyer, Partchev, Kröhne, Nagengast & Fiege, in Druck.

[2] Der Begriff *Populationsparameter* bezieht sich hier auf Parameter, welche die Verteilung von Zufallsvariablen beschreiben.

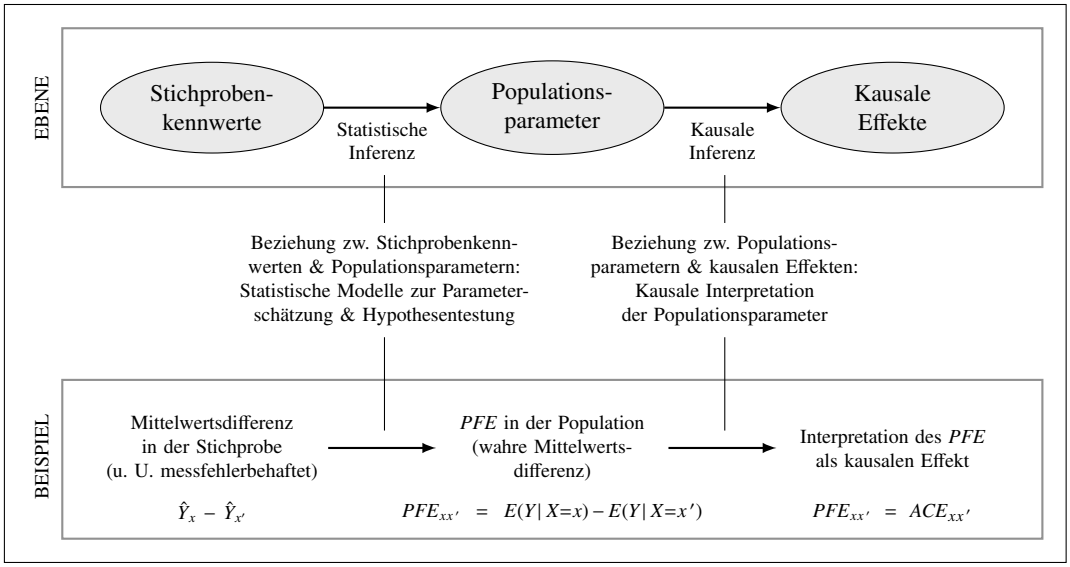

Abbildung 3.1: Schematische Darstellung verschiedener Ebenen der empirischen Kausalforschung

x' einer Stichprobe auf die wahren Mittelwertsunterschiede $E(Y|X{=}x) - E(Y|X{=}x')$ in der Population. Letztere werden auch als *Prima-Facie-Effekte* (*PFE*; Holland, 1986; vgl. Abschnitt 3.4) bezeichnet. Von der statistischen Inferenz ist die *kausale Inferenz* abzugrenzen. Diese bezieht sich auf die Inferenz von Populationsparametern auf kausale Effekte. Hier geht es also bspw. um die Frage, ob der *PFE* dem durchschnittlichen kausalen Effekt, d. h. dem *ACE* (vgl. Abschnitt 3.3), entspricht. Abbildung 3.1 zeigt ein vereinfachtes Schema der verschiedenen Ebenen empirischer Kausalforschung. Die Theorie kausaler Effekte bezieht sich insbesondere auf die kausale Inferenz.

Des Weiteren liegt der Fokus der Theorie kausaler Effekte nicht auf der Erforschung diverser Ursachen eines bestimmten Effekts, sondern auf der Identifikation des Effekts einer konkreten Ursache. Es geht also – in der Terminologie von Holland (1986) – nicht um die Analyse verschiedener „causes of effects“, sondern um die Bestimmung bzw. Quantifizierung der „effects of causes“. Die erwähnten Ursachen werden im Folgenden als *Treatment* bezeichnet. Der Treatment-Begriff ist zunächst ganz allgemein zu verstehen: Es kann sich dabei um eine medizinische, psychologische oder pädagogische Intervention handeln oder aber um das Vorhandensein anderer interessierender Wirkfaktoren. Im Kontext der Schulleistungsforschung lässt sich bspw. der Unterricht in einer bestimmten Klasse als konkretes Treatment auffassen, dessen Wirkung auf die

Schulleistung von Schülern von Interesse ist.

Zusammenfassend lässt sich das primäre Ziel wissenschaftlicher Kausalanalysen – im Rahmen einer allgemeinen Theorie kausaler Effekte – beschreiben als „ ... the investigation of selected effects of a particular cause, rather than the search for all possible causes of a particular outcome along with the comprehensive estimation of all their relative effects“ (Morgan & Harding, 2006, S. 3). Diese Perspektive wird auch im vorliegenden Kontext der Schulleistungsuntersuchungen eingenommen: Ziel ist eine Analyse von Schul- bzw. Unterrichtseffekten auf die Testleistung von Schülern und nicht die Untersuchung der multiplen Ursachen bzw. deren relativer Effekte auf die Testwerte. Die Tatsache, dass die Outcomes – hier die Testwerte der Schüler – stets durch verschiedene Faktoren beeinflusst sind (sog. *multiple Determiniertheit*), bleibt dabei jedoch keinesfalls unberücksichtigt, wie die nachfolgende Darstellung zeigt.

3.2 Terminologie und Grundkonzepte

Die Theorie kausaler Effekte basiert auf der folgenden, zunächst stark vereinfachten Grundidee, die sich bereits bei Mill (1865) findet:

> Suppose an individual, or in more general terms, an observational unit, could be treated by condition 1 or it could be treated by condition 0, *everything else invariant* [Ceteris-paribus-Klausel]. If there is a difference in the outcome considered (some measure of success of the treatment), then this difference is due to the difference in the two treatment conditions. (Steyer et al., 2011, S. VI)

Diese Grundidee spiegelt sich bspw. in Rubins Definition der *Potential-Outcomes* wider (Rubin, 1974, 1977, 1978): Der Potential-Outcome $Y_x(u)$ einer Person u in Treatment-Bedingung x ist der Wert der Outcome-Variable Y, der beobachtet werden würde, wenn diese Person der Treatment-Bedingung x ausgesetzt wird. Die Differenz der Potential-Outcomes $Y_x(u) - Y_{x'}(u)$ aus zwei verschiedenen Treatment-Bedingungen x und x' bildet dann die individuellen kausalen Effekte. Dieser Definition liegt eine deterministische Sichtweise zugrunde (*deterministic outcome assumption*, vgl. Mayer, Thoemmes, Rose & Steyer, 2011), denn der Potential-Outcome einer Person wird ausschließlich durch das Vorhandensein des jeweiligen Treatments festgelegt. Das bedeutet, dass der Potential-Outcome einer Person u den Wert $Y_x(u)$ annimmt, falls sie der Treatment-

Bedingung x ausgesetzt wird. Wird diese Person u hingegen der Treatment-Bedingung x' zugewiesen, so nimmt ihr Potential-Outcome den Wert $Y_{x'}(u)$ an.

Multiple Determiniertheit. Der Rubinschen Sichtweise ist jedoch das Argument der *multiplen Determiniertheit* entgegenzusetzen: Der Outcome einer Person wird nicht ausschließlich durch das jeweilige Treatment festgelegt, sondern kann gleichfalls durch eine Vielzahl weiterer Einflussfaktoren beeinflusst sein. Diese stochastische, nicht-deterministische Sichtweise findet sich bereits in den Überlegungen Neymans (1923/1990) im Rahmen agrarwissenschaftlicher Forschung: Nach Neyman hat jedes Stück Ackerland (das hier der Beobachtungseinheit u entspricht) bezogen auf eine jeweils betrachtete Getreidesorte (das jeweilige Treatment $X{=}x$) einen *wahren Ertrag* (den sog. *true-yield*). Der aktuell gemessene Ertrag eines Landstücks (der Wert der Outcome-Variablen Y) ist dabei lediglich eine Realisation aus der *intraindividuellen Verteilung* der Outcome-Variablen, denn dieser Ertrag kann aufgrund des Einflusses anderer Größen – wie bspw. der Regenhäufigkeit oder der Sonnenscheindauer – schwanken. Der wahre Ertrag ist dann definiert als der Erwartungswert der Outcome-Variablen Y gegeben u und x, d. h. der Erwartungswert $E(Y|\, X{=}x, U{=}u)$ dieser intraindividuellen Verteilung. Diese Sichtweise findet sich später auch bei Steyer und Kollegen (Steyer, Gabler, von Davier, Nachtigall & Buhl, 2000; Steyer, Gabler, von Davier & Nachtigall, 2000; Steyer, Nachtigall, Wüthrich-Martone & Kraus, 2002). Weiterhin definiert Neyman den individuellen kausalen Effekt als die Differenz der wahren Erträge eines Stücks Land bezogen auf zwei verschiedene Getreidesorten.

Beide Ansätze berücksichtigen jedoch lediglich die Outcome-Variable Y, die Treatment-Variable X und die Personen-Variable U im Rahmen der Definition der kausalen Effekte: Sowohl in Rubins als auch in Neymans Konzeption wird davon ausgegangen, dass über den Einfluss von X und U hinaus keine weitere systematische Beeinflussung der Outcome-Variable Y stattfindet. Das bedeutet wiederum, dass sich diese theoretischen Ansätze nur auf solche Kovariaten, d. h. potenziell konfundierende Variablen, beziehen, die deterministische Funktionen der Personen-Variable U sind. Steyer et al. (2011) hingegen definieren die elementaren Bausteine kausaler Effekte – die sog. *True-Outcomes* (vgl. Abschnitt 3.3.1) – als Erwartungswert der Outcome-Variable Y innerhalb einer Treatment-Bedingung x gegeben *sämtlicher* Kovariaten, d. h. aller der Treatment-Variable vor- oder gleichgeordneten Variablen. Diese Kovariaten können deterministische Funktionen von U sein, müssen es aber nicht sein.

Fundamentalproblem kausaler Inferenz. Hinsichtlich der oben formulierten Grundidee (vgl. S. 30) ist ein weiterer Einwand relevant: So erscheint es zunächst problematisch, kausale Effekte als Differenz zwischen den Outcomes einer Beobachtungseinheit unter Behandlung vs. unter Nicht-Behandlung zu definieren, da man eine Person zu einem konkreten Zeitpunkt stets nur entweder behandeln *oder* eben nicht behandeln kann. Beides gleichzeitig ist de facto unmöglich. Holland (1986) bezeichnet dies als „fundamental problem of causal inference“ (S. 947). Man beachte jedoch an dieser Stelle, dass sich die Theorie lediglich auf die Definition theoretischer Größen bezieht und hier noch nicht über die konkrete Durchführung eines bestimmten Experimentaldesigns oder die Methoden empirischer Datenanalyse gesprochen wird. Ein kleines Gedankenexperiment soll dies verdeutlichen: Bevor eine Person tatsächlich einer von bspw. zwei Treatment-Bedingungen zugewiesen wird, ist deren Outcome sowohl in der einen als auch in der anderen Treatment-Bedingung vorstellbar. Dies ist auch dann möglich, wenn diese Person niemals einem der Treatments zugewiesen wird. Diese Betrachtungsweise, die als *Prä-facto-Perspektive* bezeichnet wird (vgl. Steyer et al., 2011), wird im folgenden Abschnitt noch einmal verdeutlicht, in dem ich das zu betrachtende Zufallsexperiment einführe.

3.2.1 Single-Unit-Trial

Den notwendigen und hinreichenden Hintergrund zur Definition der zentralen theoretischen Begriffe bildet ein spezielles Zufallsexperiment, der sog. *Single-Unit-Trial*. Dieses Zufallsexperiment charakterisiert das in einer Anwendung jeweils betrachtete empirische Phänomen, auf das sich die stochastischen Abhängigkeiten zwischen Ereignissen bzw. zwischen Zufallsvariablen beziehen und das im Rahmen kausaler Fragestellungen analysiert werden soll. Der Single-Unit-Trial ist jedoch nicht mit der Stichprobe gleichzusetzen, auf die statistische Modelle zur Schätzung von Parametern und Testung von Hypothesen rekurrieren. Die Unterscheidung zwischen Single-Unit-Trial und Stichprobenmodell lässt sich anhand eines einfachen Zufallsexperiments – dem Werfen eines fairen Würfels[3] – veranschaulichen: Handelt es sich um einen fairen Würfel, so beträgt die Wahrscheinlichkeit für jede der sechs Augenzahlen jeweils $\frac{1}{6}$. Diese Wahrscheinlichkeit ist bereits vor dem Würfelwurf wohldefiniert und gilt selbst dann, wenn der Würfel niemals geworfen wird (Prä-facto-Perspektive). Um jedoch die Wahrschein-

[3] Ein Würfel ist *fair*, wenn jede Augenzahl mit gleicher Wahrscheinlichkeit gewürfelt wird.

lichkeit des Werfens einer bestimmten Augenzahl zu schätzen, muss eine Stichprobe erhoben werden, die aus der n-fachen Wiederholung dieses Zufallsexperiments resultiert. Die Wahrscheinlichkeit kann dann aus der relativen Häufigkeit des Auftretens dieser Augenzahl in der Stichprobe geschätzt werden.

Der Single-Unit-Trial beschreibt also die einfache bzw. einmalige Durchführung des betrachteten Zufallsexperiments. Eine Stichprobe hingegen bezieht sich auf die n-fache Wiederholung dieses Single-Unit-Trials, wobei n die Anzahl der Beobachtungen bzw. den Stichprobenumfang darstellt. Daher ermöglicht der Single-Unit-Trial auch keine Behandlung von Fragen der Parameterschätzung und Hypothesentestung. Er ist jedoch der hinreichende Hintergrund zur Definition kausaler Effekte aus einer Prä-facto-Perspektive sowie der Explikation der Bedingungen zur Berechnung (*Identifikation*) dieser theoretischen Größen mittels empirisch schätzbarer Parameter.

Im Kontext von Beobachtungsstudien – und insbesondere auch der hier betrachteten Schulleistungsstudien wie bspw. Vergleichsarbeiten – ist in der Regel der folgende Single-Unit-Trial[4] relevant:

(1) Zunächst wird eine Beobachtungseinheit u (bspw. eine Person) aus einer Menge von Beobachtungseinheiten, welche häufig als *Population* bezeichnet wird, gezogen und

(2) ihre Ausprägungen $z_1, \ldots, z_Q$ auf den Kovariaten $Z_1, \ldots, Z_Q$ (mit $Q \geq 1$) beobachtet.

(3) Diese Beobachtungseinheit wird anschließend einer von mehreren Treatment-Bedingungen (repräsentiert durch den Wert x der Treatment-Variable X) zugewiesen bzw. ihre Zuweisung wird beobachtet.

(4) Schließlich wird die numerische Ausprägung y dieser Beobachtungseinheit auf einer interessierenden Outcome-Variable Y erhoben.

[4] An dieser Stelle sei darauf hingewiesen, dass in Abhängigkeit von dem Design der Untersuchung bzw. der inhaltlichen Fragestellung verschiedene Single-Unit-Trials betrachtet werden können. Diese sollen jedoch hier nicht weiter ausgeführt werden, da dies über das Ziel der vorliegenden Arbeit hinausgeht und das hier beschriebene Zufallsexperiment im Kontext von Vergleichsarbeiten hinreichend ist. Den darüber hinaus interessierten Leser möchte ich auf die detaillierte Darstellung weiterer Single-Unit-Trials bei Steyer et al. (2011, Kapitel 2) verweisen.

Beobachtungseinheit

Der erste Schritt (1) im Rahmen des Single-Unit-Trials besteht aus der Ziehung einer Beobachtungseinheit (synonym auch *Unit*) u. In der Regel handelt es sich bei den Beobachtungseinheiten um Personen (im Kontext von Schulleistungsuntersuchungen also bspw. um einzelne Schüler einer Klasse). Eine Unit kann aber auch eine Klasse, eine Schule oder auch ein Bundesland sein. Im Rahmen dieser Arbeit beziehe ich mich in der Regel auf die Ziehung einzelner Schüler[5]. Die Variable U bezeichnet dann die Personen-Variable.

Kovariaten

In Schritt (2) werden die Ausprägungen auf potenziell konfundierenden Variablen, den Kovariaten, $Z_1, \ldots, Z_Q$ erhoben. Im Folgenden verwende ich $\boldsymbol{Z} = (Z_1, \ldots, Z_Q)$, um den Vektor dieser potenziell $Q-$dimensionalen Kovariaten zu bezeichnen. Eine Kovariate kann z. B. eine Eigenschaft der Person sein wie das Geschlecht, der sozioökonomische Status (SES) oder die Muttersprache. Sie kann jedoch auch eine fehlerbehaftete Messung einer Eigenschaft der Person sein. Ein Beispiel hierfür sind die Werte aus dem *Kognitiven Fähigkeitstest* (KFT; Heller & Perleth, 2000), mit dem die kognitiven Grundfähigkeiten von Schülern erfasst werden.

Kovariaten zeichnen sich allgemein dadurch aus, dass sie dem Treatment zeitlich vor- oder gleichgeordnet sind. Somit hat das Treatment *per definitionem* keinen Einfluss auf die Kovariaten. Dies unterscheidet Kovariaten von Mediatorvariablen (sog. *intermediate variables*), die durch ein Treatment beeinflusst werden und anschließend wiederum die Outcome-Variable Y beeinflussen (vgl. Steyer et al., 2011; Mayer et al., 2011). Im Kontext der Schulleistungsforschung kommen also nur solche Variablen als Kovariaten in Betracht, die entweder *vor (oder simultan zum)* Beginn des Unterrichts, dessen kausaler Effekt von Interesse ist, erhoben wurden *oder* die nicht durch das Treatment veränderbar sind. Letzteres trifft z. B. auf die Variable Geschlecht zu, die durch den Unterricht in einer Schule oder Klasse nicht beeinflusst werden kann.

Im Gegensatz zur zeitlichen Ordnung werden bezüglich der Art der Kovariaten keinerlei Einschränkungen gemacht: Kovariaten können ein- oder mehrdimensional, qualitativ und/oder quantitativ, manifest oder latent sein. Betrachtet man also bspw. eine mehrdimensionale Kovariate $\boldsymbol{Z}$ – bestehend aus mehreren eindimensionalen Kovariaten

[5] Andere Fälle werden explizit gekennzeichnet.

$(Z_1, \ldots, Z_Q)$ – kann diese sowohl aus qualitativen als auch aus quantitativen Kovariaten zusammengesetzt sein. Die in diesem Kapitel beschriebenen theoretischen Aspekte von Kovariaten sind notwendig, jedoch nicht hinreichend zur Beantwortung der Frage, welche konkreten Kovariaten in den empirischen Analysen berücksichtigt werden sollten. Auf diese Frage gehe ich in den nachfolgenden Kapiteln der vorliegenden Arbeit ein.

Treatment-Variable

Anschließend – in Schritt (3) – erfolgt die Zuweisung[6] der Beobachtungseinheit zu einer von mindestens zwei verschiedenen Treatment-Bedingungen. Im einfachsten Fall werden nur zwei Treatment-Bedingungen betrachtet: z. B. eine Behandlungs- und eine Kontrollbedingung ohne Behandlung. Diese Bedingungen sind die möglichen Werte der Treatment-Variable X. Es kann jedoch nicht nur zwei, sondern allgemein $J + 1$ Treatment-Bedingungen mit den Werten $0, 1, \ldots, J$ geben. Die Wahrscheinlichkeit der Treatment-Zuweisung (oder auch *Behandlungswahrscheinlichkeit*) $P(X{=}x)$ kann dabei für alle Personen gleich und bekannt sein wie bspw. im Rahmen eines randomisierten Experiments. $P(X{=}x)$ kann jedoch auch von den Eigenschaften der Personen abhängen und unbekannt sein. Letzteres betrifft z. B. Selbstselektionsprozesse, bei denen sich die Beobachtungseinheiten selbst einer der Treatment-Bedingungen zuordnen.

Outcome-Variable

Letztlich wird in Schritt (4) die numerische Ausprägung y der Beobachtungseinheit auf einer interessierenden Outcome-Variable Y, die dem Treatment zeitlich nachgeordnet ist, erhoben. Wie bei den Kovariaten $\boldsymbol{Z}$ kann es sich auch bei Y um eine potenziell multivariate Outcome-Variable handeln. Die Betrachtungen im Rahmen der vorliegenden Arbeit beschränken sich jedoch auf den univariaten Fall. Die Outcomes von Interesse sind im Folgenden die erreichten Testwerte der Schüler in den Vergleichsarbeiten (vgl. Kapitel 4).

Die im Rahmen dieses Single-Unit-Trials betrachteten Zufallsvariablen U, $\boldsymbol{Z}$, X und Y haben eine gemeinsame Verteilung auf dem zugrunde liegenden *Wahrscheinlich-*

[6] In der vorliegenden Arbeit nehme ich an, dass die Treatment-Zuweisung einer Person auch zu einer tatsächlichen Teilnahme bzw. Exposition führt, d. h., es wird ausschließlich der Fall perfekter Compliance betrachtet. Diese Annahme ist vor dem Hintergrund einer allgemeinen Schulpflicht als plausibel zu erachten.

keitsraum (vgl. Abschnitt 3.2.2), welcher das betrachtete Zufallsexperiment formal repräsentiert. Jede Kombination von Beobachtungseinheit, Ausprägung der Kovariate, Treatment-Bedingung und Wert der Outcome-Variable ist ein mögliches Ergebnis eines solchen Single-Unit-Trials, welches sich mit einer bestimmten Wahrscheinlichkeit ereignet. Mit der Spezifikation des zu betrachtenden Zufallsexperiments steht somit gleichfalls die gemeinsame Wahrscheinlichkeitsverteilung der Variablen fest, auch wenn diese zunächst vollständig unbekannt ist.

Mit dem bisher eingeführten Single-Unit-Trial lassen sich jegliche Formen stochastischer Abhängigkeiten beschreiben. Doch wann handelt es sich dabei auch um eine potenziell *kausale* stochastische Abhängigkeit zwischen Zufallsvariablen? Dafür benötigen wir den Begriff des *Kausalitätsraumes* (Steyer et al., 2011), dessen Komponenten den Ausgangspunkt für die Definition kausaler Effekte bilden.

3.2.2 Kausalitätsraum

Im folgenden Abschnitt führe ich überblickshaft die Komponenten des Kausalitätsraumes ein, welcher den Hintergrund eines stochastischen Kausalitätsmodells – und damit auch der Definition kausaler Effekte – bildet. Die zum Verständnis notwendigen Begriffe sind der Wahrscheinlichkeitstheorie entnommen. Eine detaillierte Darstellung der grundlegenden Konzepte der Wahrscheinlichkeits- und Maßtheorie findet sich z. B. bei Bauer (1990, 2001), Elstrodt (2009), Georgii (2007), Klenke (2008) sowie auch bei Steyer, Nagel, Partchev und Mayer (in Druck).

Ein Kausalitätsraum $\langle\, (\Omega, \mathfrak{A}, P),\ (\mathfrak{F}_t)_{t\in T}, X, Y, C_X \,\rangle$ besteht aus den folgenden Komponenten: (a) dem Wahrscheinlichkeitsraum $(\Omega, \mathfrak{A}, P)$, (b) den bereits eingeführten Zufallsvariablen X und Y, welche auf diesem Wahrscheinlichkeitsraum definiert sind, (c) der Filtration $(\mathfrak{F}_t)_{t\in T}$ von Sub-σ-Algebren von $\mathfrak{A}$ und (d) der umfassenden Kovariate C_X. Des Weiteren wird vorausgesetzt, dass die potenzielle Ursachenvariable X, deren kausale Effekte auf die Outcome-Variable Y betrachtet werden, der Outcome-Variable Y zeitlich vorgeordnet ist.

Wahrscheinlichkeitsraum

Die erste Komponente eines Kausalitätsraumes ist der *Wahrscheinlichkeitsraum* $(\Omega, \mathfrak{A}, P)$, der das betrachtete Zufallsexperiment repräsentiert. Der Wahrscheinlichkeitsraum ist notwendiger Bestandteil eines jeden stochastischen Modells – welches jedoch

nicht zwangsläufig auch kausale Abhängigkeiten beschreibt – und besteht aus den drei folgenden Komponenten (vgl. z. B. Bauer, 2001):

(1) Ω ist die Menge der möglichen Ergebnisse (kurz: Ergebnismenge), die im Rahmen des betrachteten Zufallsexperiments auftreten können.

(2) $\mathfrak{A}$ ist die Menge der möglichen Ereignisse. Diese ist eine σ-Algebra auf Ω, d. h. eine Menge von Teilmengen von Ω, welche abgeschlossen gegenüber endlichen und abzählbaren Vereinigungs- und Schnittmengenbildungen ist.

(3) P ist das Wahrscheinlichkeitsmaß, welches jedem Ereignis A in $\mathfrak{A}$ seine Wahrscheinlichkeit $P(A)$ zuordnet. Das Wahrscheinlichkeitsmaß P ist dabei über die Axiome von Kolmogoroff (1933) definiert, d. h. Standardisierung, Nichtnegativität und σ-Additivität.

Filtration

Eine weitere zentrale Voraussetzung für Kausalanalysen ist die zeitliche Ordnung der im Rahmen des Single-Unit-Trials betrachteten Ereignisse bzw. Zufallsvariablen: Die Kovariaten $\boldsymbol{Z}$ müssen Ereignisse repräsentieren, die *vor oder simultan mit* der Treatment-Variable X auftreten. Die Treatment-Variable X wiederum muss der Outcome-Variable Y vorgeordnet sein. Der Wahrscheinlichkeitsraum $(\Omega, \mathfrak{A}, P)$, welcher das betrachtete Zufallsexperiment repräsentiert, impliziert jedoch noch keine zeitliche Struktur der Zufallsvariablen U, $\boldsymbol{Z}$, X und Y. Diese zeitliche Ordnung von Zufallsvariablen bzw. der von ihnen dargestellten Ereignisse – die sog. *Vor- und Gleichgeordnetheit* – lässt sich mittels des Konzepts der *Filtration* $(\mathfrak{F}_t)_{t \in T}$ definieren[7]. Filtrationen sind elementarer Bestandteil der Theorie stochastischer Prozesse (z. B. Bauer, 2001; Øksendal, 2007).

Eine Filtration (oder auch *Filtrierung*) ist eine Familie monoton nichtfallender Sub-σ-Algebren von $\mathfrak{A}$. Die Filtration des im Rahmen dieser Arbeit betrachteten Zufallsexperiments (vgl. Abschnitt 3.2.1) ist in Abbildung 3.2 als Venn-Diagramm dargestellt. Die Ellipsen repräsentieren die einzelnen σ-Algebren. Die Zufallsvariablen U, $\boldsymbol{Z}$, X und Y werden jeweils innerhalb der σ-Algebren eingezeichnet, bezüglich derer sie als

[7] Dabei ist T eine Indexmenge, auf der die Relationen $=$, $<$ und $\leq$ definiert sind. Diese Relationen beziehen sich häufig – insbesondere auch in der vorliegenden Arbeit – auf zeitliche Verhältnisse (d. h. gleichzeitig, zeitlich vorgeordnet sowie gleichzeitig *und* zeitlich vorgeordnet). Des Weiteren gelte im Rahmen der nachfolgenden Betrachtungen $T = \{1, 2, \ldots, n\},\ n \in \mathbb{N}$.

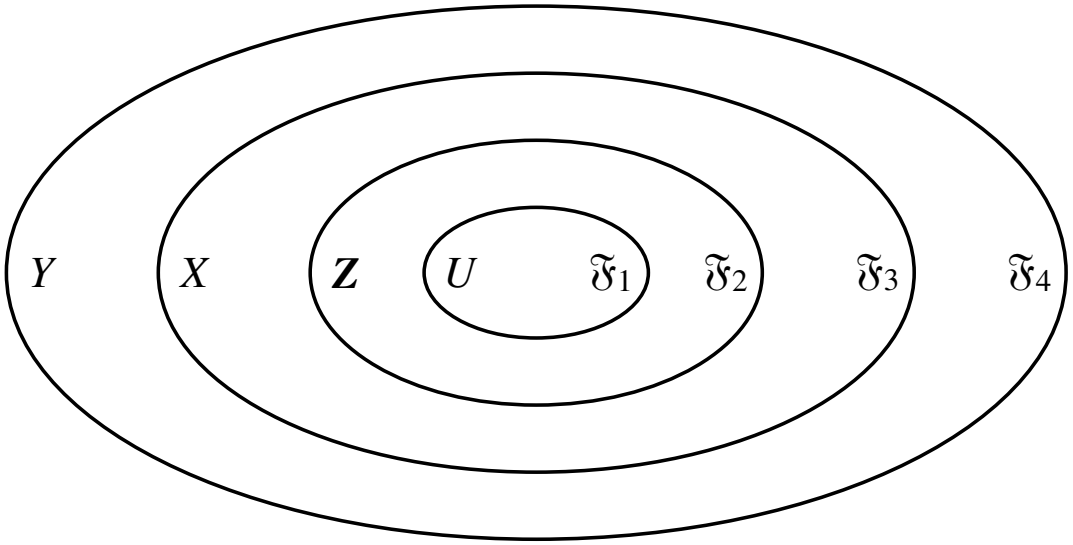

Abbildung 3.2: Venn-Diagramm der Filtration $(\mathfrak{F}_t)_{t\in T}$ mit vier σ-Algebren $(T = \{1, 2, 3, 4\})$

erstes messbar sind[8] (für eine ähnliche Form der Darstellung vgl. Nagengast, 2009). Die Menge der möglichen Ereignisse $\mathfrak{A}$ des Wahrscheinlichkeitsraumes $(\Omega, \mathfrak{A}, P)$ enthält hier also die Filtration $(\mathfrak{F}_t)_{t\in T}$ von vier σ-Algebren[9], d. h. $\mathfrak{F}_t \subset \mathfrak{A}$, $t \in T$ mit $T = \{1, 2, 3, 4\}$. Dabei sind $\mathfrak{F}_1$ und $\mathfrak{F}_2$ die σ-Algebren, welche durch alle dem Treatment zeitlich vorgeordneten Ereignisse erzeugt werden. $\mathfrak{F}_3$ ist die σ-Algebra, welche durch alle dem Treatment zeitlich vor- und gleichgeordneten Ereignisse erzeugt wird und $\mathfrak{F}_4$ ist die σ-Algebra, welche durch alle dem Treatment zeitlich vor-, gleich- und nachgeordneten Ereignisse gebildet wird. Weiterhin sei $\mathfrak{F}_1$ den drei σ-Algebren $\mathfrak{F}_2$, $\mathfrak{F}_3$ und $\mathfrak{F}_4$ vorgeordnet, $\mathfrak{F}_2$ sei $\mathfrak{F}_3$ und $\mathfrak{F}_4$ vorgeordnet und schließlich sei $\mathfrak{F}_3$ bezüglich $\mathfrak{F}_4$ vorgeordnet (für eine formale Definition des Begriffs *Vorgeordnetheit* vgl. Steyer, 1992 und Steyer et al., 2011, Kapitel 3).

Vorgeordnetheit ist auch für Zufallsvariablen definiert. Bei dieser Definition greift man auf die von den Zufallsvariablen erzeugten σ-Algebren zurück. Dabei ist $\mathfrak{F}_1$ die

[8] Hier wird eine Vereinfachung in der Darstellung vorgenommen, denn solche Variablen $\boldsymbol{Z}$, die deterministische Funktionen $f(U)$ der Personen-Variable U sind, sind bereits messbar bezüglich $\mathfrak{F}_1$. $\boldsymbol{Z}$ ist in der Abbildung 3.2 zum Zwecke der Übersichtlichkeit jedoch erst in $\mathfrak{F}_2$ eingezeichnet. Eine detaillierte Erläuterung zu Abbildung 3.2 findet sich unten im Text. Zum Prinzip der *Messbarkeit von Zufallsvariablen* verweise ich auf Bauer (2001), Georgii (2007) oder Klenke (2008).

[9] Im Rahmen eines Kausalitätsraumes bei Experimenten oder Quasi-Experimenten muss die Filtration mindestens aus drei σ-Algebren bestehen, um den einfachsten Single-Unit-Trial repräsentieren zu können (vgl. Steyer et al., 2011).

von der Personen-Variable U erzeugte σ-Algebra, wohingegen $\mathfrak{F}_2$ die Vereinigung von $\mathfrak{F}_1$ *und* der von der Kovariate $\boldsymbol{Z}$ erzeugten σ-Algebra ist. Die (potenziell multivariate) Variable $\boldsymbol{Z}$ kann dabei eine deterministische Funktion $\boldsymbol{f(U)}$ der Personen-Variablen U sein, d. h. Eigenschaften der Personen wie bspw. Geschlecht, SES oder Muttersprache. Solche Variablen $\boldsymbol{Z}$, die deterministische Funktionen von U sind, sind bereits messbar bezüglich $\mathfrak{F}_1$. Jedoch kann $\boldsymbol{Z}$ auch eine fehlerbehaftete Funktion der Personen-Variable U sein, d. h. eine fehlerbehaftete Messung einer Eigenschaft der Personen wie bspw. die *Kognitiven Grundfähigkeiten* eines Schülers. Solche latente Variablen sind jeweils per definitionem eine Funktion von U (vgl. Steyer, 1989, 2001). Bei gegebener Person u sind die Werte von $\boldsymbol{Z}$ jedoch nicht deterministisch festgelegt, sondern haben eine intraindividuelle Verteilung. In diesem Fall gilt: $\boldsymbol{Z} = \boldsymbol{f(U)} + \boldsymbol{\varepsilon}$, d. h., $\boldsymbol{Z}$ ist die Summe einer deterministischen Funktion $\boldsymbol{f(U)}$ der Personen-Variablen U (die *wahren Werte* dieser Kovariaten) und einer Messfehlervariablen $\boldsymbol{\varepsilon}$. Dabei kann $\boldsymbol{\varepsilon}$ prinzipiell einen zusätzlichen, über X und U hinausgehenden Effekt auf die Outcome-Variable Y haben. Dann sind sowohl die fehlerbehafteten Messungen $\boldsymbol{Z}$ als auch $\boldsymbol{\xi} \equiv \boldsymbol{f(U)}$ als Kovariaten zu betrachten, wobei $\boldsymbol{Z}$ eine fehlerbehaftete *manifeste* und $\boldsymbol{\xi}$ eine *latente* Kovariate ist. Die latente Variable $\boldsymbol{\xi}$ ist wiederum bereits messbar bezüglich $\mathfrak{F}_1$. Zusammenfassend ist die σ-Algebra $\mathfrak{F}_2$ so definiert, dass alle dem Treatment vorgeordneten Zufallsvariablen messbar bezüglich $\mathfrak{F}_2$ sind. Dies umfasst Funktionen von U (bspw. latente Variablen), fehlerbehaftete Messungen von Eigenschaften der Personen, aber auch messbare Funktionen beider Arten von Variablen. Weiterhin ist $\mathfrak{F}_3$ die Vereinigung von $\mathfrak{F}_2$ *und* der von der Treatment-Variable X erzeugten σ-Algebra. Schließlich ist $\mathfrak{F}_4$ die Vereinigung von $\mathfrak{F}_3$ und der von der Outcome-Variable Y erzeugten σ-Algebra.

Die umfassende Kovariate C_X

Mittels des Konzepts der Filtration lässt sich die zeitliche Ordnung der Ereignisse (bzw. Ereignismengen) definieren, die im Rahmen eines Zufallsexperiments auftreten können. Dabei ist entscheidend, dass die Ursache X den Kovariaten nicht vorgeordnet sein kann. Daher betrachten wir im Folgenden eine bestimmte σ-Algebra aus der Filtration $(\mathfrak{F}_t)_{t\in T}$, die sog. *umfassende Kovariaten-σ-Algebra* $\mathfrak{C}_X$. Hierbei handelt es sich um die größte Sub-σ-Algebra aus der Filtration $(\mathfrak{F}_t)_{t\in T}$, bezüglich der X noch nicht messbar ist. Vereinfacht ausgedrückt ist $\mathfrak{C}_X$ so definiert, dass sie alle Ereignisse enthält, welche *vor oder simultan mit* Einsetzen des Treatments X auftreten können, nicht jedoch die

Ereignisse, die durch X selbst repräsentiert werden (für eine formale und allgemeine Definition der umfassenden Kovariaten-σ-Algebra $\mathfrak{C}_X$ vgl. Steyer et al., 2011, Kapitel 3). Somit ist die umfassende Kovariaten-σ-Algebra $\mathfrak{C}_X$ stets bezüglich der jeweils betrachteten potenziellen Ursache X definiert.

Weiterhin ist die *umfassende Kovariate* C_X messbar bezüglich $\mathfrak{C}_X$ bzw. $\mathfrak{C}_X$ ist gleichzeitig die durch C_X erzeugte σ-Algebra. Die umfassende Kovariate C_X repräsentiert *alle Kovariaten*, d. h. alle der Treatment-Variable vor- oder gleichgeordneten Variablen, die im Rahmen eines konkreten Zufallsexperiments auftreten können. In vielen Anwendungen ist C_X gleichzusetzen mit der Personen-Variable U. Hierzu zählen wiederum sämtliche deterministische Funktionen $\boldsymbol{f(U)}$ der Personen-Variable U, also bspw. Eigenschaften der Person, die ohne Messfehler erhoben werden. Vereinfacht ausgedrückt bedeutet dies, dass sich die systematische Variabilität bezüglich der Outcome-Variable Y, die nicht auf das Treatment zurückzuführen ist, ausschließlich durch Unterschiede zwischen Personen erklären lässt. Gegeben eine konkrete Person u hängt der Outcome einzig von dem Vorhandensein oder Nicht-Vorhandensein des Treatments ab. Bei messfehlerbehafteten Kovariaten kann es jedoch auch innerhalb der Personen noch systematische Variabilität geben, die nicht allein auf das Treatment zurückzuführen ist. In diesem Fall gilt: $C_X = (U, \boldsymbol{Z})$. Dies bedeutet, dass auch messfehlerbehaftete Kovariaten in C_X berücksichtigt sein können, da sich C_X auf alle Ereignisse bezieht, die dem Treatment zeitlich vor- bzw. gleichgeordnet sind.

3.3 Kausale Effekte

Der Single-Unit-Trial sowie die Komponenten des Kausalitätsraumes bilden den Ausgangspunkt der im Folgenden zu definierenden theoretischen Konzepte: die *True-Outcome-Variable*, die *True-Effect-Variable* sowie die verschiedenen Arten *durchschnittlicher* und *bedingter kausaler Effekte*. Letztere repräsentieren die im Rahmen einer konkreten Anwendung interessierenden theoretischen Zielgößen. Sämtliche in diesem Abschnitt dargestellten Definitionen basieren auf Steyer et al. (2011).

Steyer et al. (2011) unterscheiden zwischen totalen, direkten und indirekten Effekten, wobei potenzielle Mediatorvariablen berücksichtigt werden können. Diese Unterscheidung ist im Rahmen der vorliegenden Arbeit nicht relevant, da ich ausschließlich die totalen Effekte betrachte. Daher verwende ich bei den folgenden Definitionen den Be-

griff *kausale Effekte*, der synonym zu den bei Steyer et al. (2011) verwendeten Terminus der *totalen Effekte* ist.

3.3.1 True-Outcome-Variable und True-Effect-Variable

Mit der umfassenden Kovariate C_X (vgl. Abschnitt 3.2.2) werden alle konfundierenden Variablen berücksichtigt, welche die Outcome-Variable Y über das Treatment hinaus beeinflussen können. Somit ist C_X der elementare Baustein für die Definition der *True-Outcome-Variable*[10] τ_x:

$$\tau_x \equiv E^{X=x}(Y|\, C_X)\ . \tag{3.1}$$

Die True-Outcome-Variable τ_x ist also definiert als die bedingte Regression[11] $E^{X=x}(Y|\, C_X)$ der Outcome-Variable Y auf die umfassende Kovariate C_X in einer Treatment-Bedingung x (mit $x = 0, 1, \ldots, J$)[12]. Die Grundidee bezüglich der Definition der True-Outcome-Variable τ_x besteht darin, auf die sog. *atomaren Strata* (engl.: *atomic strata*, vgl. Steyer et al., 2011, Kapitel 4) zu bedingen, d.h auf die Werte der umfassenden Kovariate C_X. Wir bedingen also auf die feinste Ebene, so dass es keinen über X und C_X hinausgehenden Effekt auf Y mehr geben kann. Im Gegensatz zur Outcome-Variable Y ist die True-Outcome-Variable τ_x somit bereinigt von jeglichen *Bias* (Verzerrungen), d. h. verfälschenden, konfundierenden Effekten: Durch das Bedingen auf die umfassende Kovariate C_X ist τ_x so konstruiert, dass τ_x von allen konfundierenden Effekten bereinigt ist.

Weiterhin ist die *True-Effect-Variable* $\delta_{xx'}$ von Treatment x im Vergleich zu Treat-

[10] Genauer formuliert handelt es sich hier um die *True-Outcome-Variable bezüglich totaler Effekte* (*true-outcome variables with respect to total effects*; vgl. Steyer et al., 2011, Kapitel 4). Da ich mich – wie bereits erwähnt – im Rahmen der vorliegenden Arbeit ausschließlich auf totale Effekte beziehe, wird in den folgenden Ausführungen der vereinfachte Ausdruck *True-Outcome-Variable* verwendet. Für eine allgemeinere Darstellung, im Rahmen derer auch direkte und indirekte Effekte betrachtet werden können, verweise ich auf Steyer et al. (2011) sowie Mayer et al. (2011). Zugunsten der Lesbarkeit werde ich nachfolgend auf ähnliche Anpassungen bezüglich der im Weiteren verwendeten Begriffe nicht mehr zusätzlich hinweisen.

[11] Die bedingte Regression $E^{X=x}(Y|\, C_X)$ ist die Regression von Y auf C_X bezüglich des bedingten Wahrscheinlichkeitsmaßes $P^{X=x}$, das definiert ist durch $P^{X=x}(A) = P(A \mid X{=}x)$ für alle $A \in \mathfrak{A}$ (vgl. Steyer et al., 2011, Kapitel 4).

[12] Im Rahmen der Definition der True-Outcome-Variablen τ_x wird zusätzlich die Annahme der sog. *P-Uniqueness* gemacht. Diese Annahme stellt sicher, dass die True-Outcome-Variablen τ_x wohldefiniert sind. Sind alle True-Outcome-Variablen τ_x *P-unique*, dann können auch deren Differenzen $\tau_x - \tau_{x'}$ sowie deren unbedingte und bedingte Erwartungswerte betrachtet werden. Für eine detaillierte Darstellung der *P-Uniqueness*-Annahme verweise ich auf Steyer et al., 2011, Kapitel 4.

ment x' auf die Outcome-Variable Y wie folgt definiert:

$$\delta_{xx'} \equiv \tau_x - \tau_{x'}\,, \tag{3.2}$$

d. h. als Differenz der True-Outcome-Variablen τ_x und $\tau_{x'}$. Die Werte der True-Outcome-Variablen $\delta_{xx'}$ sind die Effekte von Treatment x vs. Treatment x' auf die Outcome-Variable Y bei gegebener Ausprägung von C_X. Die Effekte einer Treatment-Bedingung x sind also stets relativ zu einer konkreten Vergleichsbedingung x' definiert.

3.3.2 Durchschnittliche und bedingte kausale Effekte

Basierend auf den zentralen Konzepten der True-Outcome-Variable τ_x sowie der True-Effect-Variable $\delta_{xx'}$, welche sich auf den Single-Unit-Trial beziehen, lassen sich nun die verschiedenen Arten durchschnittlicher und bedingter kausaler Effekte definieren.

Durchschnittlicher kausaler Effekt

Der *durchschnittliche kausale Effekt* (*average causal effect* oder $ACE_{xx'}$) von Treatment x im Vergleich zu Treatment x' auf die Outcome-Variable Y ist definiert als Erwartungswert der True-Effect-Variable $\delta_{xx'}$:

$$ACE_{xx'} \equiv E(\delta_{xx'})\,. \tag{3.3}$$

Weiterhin folgt aus der Definition $\delta_{xx'} \equiv \tau_x - \tau_{x'}$ (vgl. Gleichung 3.2) sofort: $ACE_{xx'} = E(\tau_x) - E(\tau_{x'})$. Der $ACE_{xx'}$ ist der durchschnittliche kausale Effekt von Treatment x vs. Treatment x' für die betrachtete Gesamtpopulation. Bezug nehmend auf Schulleistungsuntersuchungen in Deutschland könnte bspw. von Interesse sein, ob eine Verlängerung der Grundschulzeit auf sechs Jahre – im Vergleich zu vier Jahren – einen positiven Effekt auf die Lernentwicklung von Schülern *im Allgemeinen*, d. h. der Schüler aller deutschen Schulen, hat.

Bedingte kausale Effekte

Bedingte kausale Effekte (*conditional causal effects* oder $CCE_{xx';V=v}$) sind allgemein definiert als

$$CCE_{xx';V=v} \equiv E(\delta_{xx'} \mid V{=}v)\,, \tag{3.4}$$

wobei v die Werte einer (zunächst beliebigen) Zufallsvariable V auf dem Wahrscheinlichkeitsraum $(\Omega, \mathfrak{A}, P)$ des betrachteten Single-Unit-Trials sind. Das gemeinsame Prinzip der bedingten kausalen Effekte besteht somit darin, den $(V{=}v)$-bedingten Erwartungswert der True-Effect-Variablen $\delta_{xx'}$ zu betrachten. Nachfolgend behandele ich verschiedene Beispiele für V, wobei es sich um die Personen-Variable U, die Kovariaten $\boldsymbol{Z}$ oder auch die Treatment-Variable X handeln kann.

Weiterhin sind die $(V{=}v)$-bedingten kausalen Effekte $E(\delta_{xx'} \mid V{=}v)$ die Werte der *V-bedingten kausalen Effektfunktion* $E(\delta_{xx'} \mid V)$. Für diese gilt:

$$E[\, E(\delta_{xx'} \mid V)\,] = E(\delta_{xx'})\,, \tag{3.5}$$

d. h., der Erwartungswert der V-bedingten kausalen Effektfunktion ist gleich dem durchschnittlichen kausalen Effekt $ACE_{xx'}$.

Individueller kausaler Effekt. Einen Spezialfall der $(V{=}v)$-bedingten Effekte stellen die *individuellen kausalen Effekte* (*individual causal effects* oder $CCE_{xx';\,U=u}$) dar. Hierbei bedingen wir auf die Personen-Variable U. Der individuelle kausale Effekt eines Treatments x im Vergleich zu einem Treatment x' für eine Person u ist wie folgt definiert:

$$CCE_{xx';\,U=u} \equiv E(\delta_{xx'} \mid U{=}u)\,. \tag{3.6}$$

Auch hier folgt aus der Definition $\delta_{xx'} \equiv \tau_x - \tau_{x'}$ (vgl. Gleichung 3.2), dass $CCE_{xx';\,U=u} = E(\tau_x|U{=}u) - E(\tau_{x'}|U{=}u)$. Der individuelle kausale Effekt $CCE_{xx';\,U=u}$ ist demnach die Differenz zwischen den True-Outcomes gegeben Treatment x und gegeben Treatment x' innerhalb einer Person u. Angewandt auf den Kontext von Schulleistungsuntersuchungen ist bspw. der individuelle kausale Effekt eines Unterrichts in einer Klasse x im Vergleich zum Unterricht in einer anderen Klasse x' für einen bestimmten Schüler u die Differenz der zu erwartenden Testleistungen, die dieser Schüler unter den jeweiligen Unterrichtsformen erzielen würde. Die Verwendung des Konjunktivs II in dem beschriebenen Beispiel soll die *präfaktische* Sichtweise im Rahmen stochastischer Modelle, denen auch die Theorie kausaler Effekte zuzuordnen ist, deutlich machen. Die True-Outcomes charakterisieren Eigenschaften des zugrunde liegenden Zufallsexperiments (vgl. Abschnitt 3.2.1) und können nur aus einer Prä-facto-Perspektive, d. h. aus einer Perspektive, *bevor* das Zufallsexperiment durchgeführt wird, adäquat interpretiert werden. Aus dieser Sicht lässt sich der Erwartungswert des Schülers u bezüglich seiner

Testleistung sowohl unter Treatment x als auch unter Treatment x' betrachten – im Sinne einer hypothetischen Eigenschaft des Schülers – auch wenn der jeweilige Schüler u weder Treatment x noch Treatment x' ausgesetzt wird.

Der individuelle kausale Effekt stellt also eine antizipierbare Eigenschaft der Beobachtungseinheit dar, auch wenn das betrachtete Zufallsexperiment niemals durchgeführt wird. Damit ist der individuelle kausale Effekt – trotz des Fundamentalproblems kausaler Inferenz – ein wohldefinierter Begriff. Auch im Alltag antizipieren wir die Effekte bestimmten Handelns und vermeiden es, wenn das antizipierte Risiko als zu hoch eingeschätzt wird. So ist bspw. der Konsum von Gurken im Mai und Juni des Jahres 2011 drastisch gesunken, nachdem die Quelle der gefährlichen EHEC-Infektion bei diesem Gemüse vermutet wurde. Das antizipierte Risiko einer Erkrankung führte bei vielen Deutschen dazu, auf den Konsum von Gurken zu verzichten.

Da die individuellen kausalen Effekte in empirischen Anwendungen in der Regel nicht identifiziert werden können, liegt das Hauptaugenmerk kausaler Effektanalysen auf den durchschnittlichen kausalen Effekten oder auch den beiden nachfolgend beschriebenen bedingten kausalen Effekten: der bedingte kausale Effekt bei gegebenem Wert einer Kovariate bzw. der bedingte kausale Effekt bei gegebenem Wert der Treatment-Variable. Unter Gültigkeit bestimmter Annahmen (vgl. Abschnitt 3.4) lassen sich diese aus Stichprobendaten im Rahmen empirischer Anwendungen schätzen.

Bedingter kausaler Effekt bei gegebenem Wert einer Kovariate. Interessiert man sich weniger für den $ACE_{xx'}$ in der Gesamtpopulation, sondern vielmehr für den durchschnittlichen kausalen Effekt in spezifischen Subpopulationen, die durch die Werte z einer (ein- oder auch mehrdimensionalen) Kovariaten $\boldsymbol{Z}$ charakterisiert werden, sind die ($\boldsymbol{Z}$=z)-*bedingten kausalen Effekte* informativer:

$$CCE_{xx';\boldsymbol{Z}=z} \equiv E(\delta_{xx'} \mid \boldsymbol{Z}=z)\,. \tag{3.7}$$

Demnach ist der $CCE_{xx';\boldsymbol{Z}=z}$ der durchschnittliche kausale Effekt von Treatment x im Vergleich zu Treatment x' auf die Outcome-Variable Y bei gegebenem Wert z einer Kovariaten $\boldsymbol{Z}$. Des Weiteren folgt aus der Definition $\delta_{xx'} \equiv \tau_x - \tau_{x'}$ (vgl. Gleichung 3.2) sofort: $CCE_{xx';\boldsymbol{Z}=z} = E(\tau_x \mid \boldsymbol{Z}=z) - E(\tau_{x'} \mid \boldsymbol{Z}=z)$. Diese ($\boldsymbol{Z}$=$z$)-bedingten kausalen Effekte spielen bspw. dann eine Rolle, wenn man subgruppenspezifische Effekte beschreiben

möchte. Im Kontext von Schulleistungsuntersuchungen könnte also von Interesse sein, ob sich differentielle Effekte einer neuen Unterrichtsform in der Gruppe leistungsstarker Schüler vs. der Gruppe leistungsschwacher Schüler zeigen. In diesem Beispiel ist $\boldsymbol{Z}$ die (dichotomisierte) Leistung der Schüler vor der Einführung der neuen Unterrichtsmethode.

Bedingter kausaler Effekt bei gegebenem Wert der Treatment-Variable. Neben den bereits erwähnten *ACE*- und *CCE*-Arten gibt es eine weitere Effektdefinition, die im Rahmen der vorliegenden Arbeit von besonderem Interesse ist: der durchschnittliche kausale Effekt $CCE_{xx';X=x^*}$ eines Treatments x im Vergleich zu Treatment x' gegeben Treatment x^*. Formal ist der $(X{=}x^*)$-*bedingte kausale Effekt* wie folgt definiert:

$$CCE_{xx';X=x^*} \equiv E(\delta_{xx'} \mid X{=}x^*)\,. \tag{3.8}$$

Auch hier folgt aus $\delta_{xx'} \equiv \tau_x - \tau_{x'}$ (vgl. Gleichung 3.2) wiederum sofort: $CCE_{xx';X=x^*} = E(\tau_x|X{=}x^*) - E(\tau_{x'}|X{=}x^*)$. Für $x^* = x$ ist der $CCE_{xx';X=x^*}$ der *average causal effect on the treated* (vgl. z. B. Rubin, 1977), den ich im Verlauf dieser Arbeit auch mit *ACE on the treated* abkürze[13].

Die inhaltliche Interpretation dieses kausalen Effekts erscheint im Vergleich zu den $(\boldsymbol{Z}{=}z)$-bedingten kausalen Effekten zunächst weniger intuitiv. Die $(X{=}x^*)$-bedingten kausalen Effekte enthalten nur dann spezifischere Informationen als der (unbedingte) durchschnittliche kausale Effekt, wenn die bedingten Erwartungswerte der True-Outcome-Variable τ_x von X abhängen, d. h. wenn gilt: $E(\tau_x \mid X) \neq E(\tau_x)$. Andernfalls sind beide theoretische Größen gleich; es gilt somit: $CCE_{xx';X=x^*} = ACE_{xx'}$ für alle Werte x^* von X. Anschaulich ausgedrückt bedeutet dies: Haben z. B. diejenigen Personen, die systematisch mehr von dem betrachteten Treatment profitieren würden, größere (oder auch kleinere) Wahrscheinlichkeiten, diesem Treatment zugewiesen zu werden, als diejenigen, die vergleichsweise weniger davon profitieren würden, dann gilt: $CCE_{xx';X=x^*} \neq ACE_{xx'}$. Dies kann dann der Fall sein, wenn die Zuordnung der Untersuchungseinheiten zu den Treatment-Bedingungen nicht randomisiert, sondern systematisch (bspw. durch Selbstselektion) erfolgt. Da die Zuweisung zu den

[13] In der Literatur wird dieser Effekt auch als *effect of treatment on the treated* (vgl. Geneletti & Dawid, 2011; Heckman & Robb, 1985; Hernán & Robins, 2006; Shpitser & Pearl, 2009; Stuart, 2004) oder *average treatment effect on the treated* (vgl. Hotz, Imbens & Klerman, 2006; Morgan & Winship, 2007) bezeichnet.

Treatment-Gruppen im Rahmen von Beobachtungsstudien, denen auch Schulleistungsuntersuchungen und Vergleichsarbeiten zuzuordnen sind, nicht der Kontrolle der Durchführenden unterliegt, ist hier von derartigen Selektionseffekten auszugehen, denn die Schüler werden einzelnen Klassen oder auch Schulen natürlich nicht per Randomisierung zugeordnet. So werden sich bspw. die Schüler zweier Klassen x und x' an zwei verschiedenen Schulen bezüglich des sozioökonomischen Status, welcher neben dem Unterricht ebenfalls die Leistung der Schüler beeinflusst, unterscheiden. Hier könnte demnach der Effekt $CCE_{xx';X=x}$ des Unterrichts in Klasse x auf die Schüler dieser Klasse anders ausfallen als der Effekt $CCE_{xx';X=x'}$ des Unterrichts in Klasse x auf die Schüler der *anderen* Klasse x' haben würde. An diesem Beispiel wird weiterhin deutlich, dass im Rahmen von Schulleistungsuntersuchungen weniger der durchschnittliche kausale Effekt $ACE_{xx'}$ in der Gesamtpopulation, die sich hier aus den Schülern beider Klassen zusammensetzt, von inhaltlichem Interesse ist, sondern vielmehr der bedingte kausale Effekt $CCE_{xx';X=x}$ bezogen auf die Schüler der betrachteten Klasse x[14]. Der Lehrer einer Klasse wird in aller Regel auf die Optimierung der Unterrichtseffekte bezüglich der jeweils *eigenen* Schülerschaft abzielen und nicht bezüglich Schüler im Allgemeinen, die zu großen Teilen gar nicht als Schüler seiner Klasse in Frage kommen.

3.3.3 Effektparametrisierung kausaler Effekte

Bisher haben wir paarweise Vergleiche zwischen jeweils zwei unterschiedlichen Treatment-Bedingungen x und x' im Rahmen der Definition kausaler Effekte betrachtet. Steyer et al. (2011, Kapitel 5) bezeichnen dies auch als *Differenzparametrisierung*. Für eine Vielzahl von Forschungsfragen ist diese Betrachtung vollkommen ausreichend: Häufig geht es um den Vergleich einer Behandlungsgruppe mit einer Kontrollgruppe, die keiner oder auch einer alternativen Behandlungsform ausgesetzt wurde. Sobald jedoch mehrere Treatment-Stufen Gegenstand der Untersuchung sind, stellt sich die Frage, welche Treatment-Bedingung jeweils als Referenz herangezogen werden sollte. Das ist v. a. dann problematisch, wenn es keine explizite Kontrollbedingung gibt. Genau dies ist der Fall in Schulleistungsuntersuchungen wie bspw. Vergleichsarbeiten. Hier werden viele Treatment-Bedingungen, welche jeweils den Unterricht in den verschiedenen Klassen bzw. Schulen repräsentieren, betrachtet. Eine mögliche Vorgehens-

[14] Entsprechend ist für die Schüler der Klasse x' der bedingte kausale Effekt $CCE_{x'x;X=x'}$ von primärem Interesse.

weise besteht nun darin, die True-Outcome-Variable gegeben Treatment x nicht mit der True-Outcome-Variable eines anderen Treatments x' zu vergleichen, sondern mit dem gleichgewichteten Mittelwert der True-Outcome-Variablen aller $J + 1$ Treatment-Bedingungen. Dieses Vorgehen bezeichnen Steyer et al. (2011, Kapitel 5) auch als *Effektparametrisierung* kausaler Effekte. Im Folgenden verwende ich die Notation nach Steyer et al. (2011): Dabei entfällt der Index x' für die zweite Treatment-Bedingung, da die Referenz stets der Mittelwert der True-Outcome-Variablen aller $J + 1$ Werte von X ist. Weiterhin bezeichne ich die nachfolgend zu definierenden theoretischen Größen als (durchschnittliche und bedingte) *kausale Effekte von x*. Die True-Effect-Variable von x ist dann wie folgt definiert:

$$\delta_x \equiv \tau_x - \frac{1}{J+1} \sum_{x'=0}^{J} \tau_{x'} \,. \tag{3.9}$$

Der durchschnittliche kausale Effekt von x ist wiederum der Erwartungswert dieser True-Effect-Variable:

$$ACE_x \equiv E(\delta_x) \,. \tag{3.10}$$

Auf individueller Ebene liegt dann folgende alternative Definition des individuellen kausalen Effekts von x nahe:

$$CCE_{x;\,U=u} \equiv E(\delta_x \mid U{=}u) \,. \tag{3.11}$$

Für den ($\boldsymbol{Z}{=}\boldsymbol{z}$)-bedingten kausalen Effekt von Treatment x verglichen mit dem Durchschnitt aller Treatment-Bedingungen bei gegebenem Wert $\boldsymbol{z}$ der Kovariaten $\boldsymbol{Z}$, also den ($\boldsymbol{Z}{=}\boldsymbol{z}$)-bedingten kausalen Effekt von x, ergibt sich demnach:

$$CCE_{x;\,\boldsymbol{Z}=\boldsymbol{z}} \equiv E(\delta_x \mid \boldsymbol{Z}{=}\boldsymbol{z}) \,. \tag{3.12}$$

Entsprechend ist der durchschnittliche kausale Effekt von Treatment x verglichen mit dem Durchschnitt aller Treatment-Bedingungen gegeben Treatment-Bedingung x^*, d. h. der ($X{=}x^*$)-bedingte kausale Effekt von x, definiert als:

$$CCE_{x;\,X=x^*} \equiv E(\delta_x \mid X{=}x^*) \,. \tag{3.13}$$

Tabelle 3.1: Übersicht der wichtigsten kausalen Effekte in der Differenz- und Effektparametrisierung

Definition	Bedeutung	Kurzform
Differenzparametrisierung:		
$\tau_x - \tau_{x'}$	True-Effect-Variable x vs. x'	$\delta_{xx'}$
$E(\delta_{xx'})$	Durchschnittlicher kausaler Effekt von x vs. x'	$ACE_{xx'}$
$E(\delta_{xx'} \mid U{=}u)$	Individueller kausaler Effekt von x vs. x'	$CCE_{xx';\,U=u}$
$E(\delta_{xx'} \mid \boldsymbol{Z}{=}\boldsymbol{z})$	$(\boldsymbol{Z}{=}\boldsymbol{z})$-bedingter kausaler Effekt von x vs. x'	$CCE_{xx';\,\boldsymbol{Z}=\boldsymbol{z}}$
$E(\delta_{xx'} \mid X{=}x^*)$	$(X{=}x^*)$-bedingter kausaler Effekt von x vs. x'	$CCE_{xx';\,X=x^*}$
Effektparametrisierung:		
$\tau_x - \frac{1}{J+1}\sum_{x'=0}^{J}\tau_{x'}$	True-Effect-Variable von x	δ_x
$E(\delta_x)$	Durchschnittlicher kausaler Effekt von x	ACE_x
$E(\delta_x \mid U{=}u)$	Individueller kausaler Effekt von x	$CCE_{x;\,U=u}$
$E(\delta_x \mid \boldsymbol{Z}{=}\boldsymbol{z})$	$(\boldsymbol{Z}{=}\boldsymbol{z})$-bedingter kausaler Effekt von x	$CCE_{x;\,\boldsymbol{Z}=\boldsymbol{z}}$
$E(\delta_x \mid X{=}x^*)$	$(X{=}x^*)$-bedingter kausaler Effekt von x	$CCE_{x;\,X=x^*}$

Beim $CCE_{x;\,X=x^*}$ handelt es sich um die Effektparametrisierung des *ACE on the treated*. An dieser Stelle sei noch einmal betont, dass in den Gleichungen 3.9 bis 3.13 eine Gleichgewichtung mit $1/(J+1)$ bei der Durchschnittsbildung über alle $J+1$ Treatment-Stufen erfolgt. Designspezifische Aspekte hingegen wie bspw. die Behandlungswahrscheinlichkeiten $P(X{=}x)$ fließen nicht in die Definition der theoretischen Größen ein (vgl. Fiege, 2007; Steyer et al., 2011). Bei der *Definition* des kausalen Effekts einer Behandlungsform (im Vergleich zu verschiedenen anderen Treatments) sollten Design-Fragen keine Rolle spielen, da diese für die Bewertung des Effekts der Behandlung *per se* keine Bedeutung haben. Des Weiteren beziehen sich auch die Effektdefinitionen in Gleichungen 3.9 bis 3.13 auf den Single-Unit-Trial (vgl. Abschnitt 3.2.1). Design-Aspekte werden jedoch erst dann relevant, wenn man zur Betrachtung empirischer Daten – bspw. im Rahmen von Zwischengruppendesigns, die in der Regel auch bei Schulleistungsuntersuchungen realisiert sind – übergeht.

Tabelle 3.1 fasst die wichtigsten kausalen Effekte, deren Beschreibung sowie die im weiteren Verlauf dieser Arbeit verwendeten Kurzformen zusammen. Dabei ist sowohl die Differenz- als auch die Effektparametrisierung aufgeführt.

3.4 Identifikation kausaler Effekte

Bisher wurden lediglich die *Definitionen* von theoretischen Größen betrachtet. Diese ermöglichen jedoch noch keine Aussagen im Rahmen empirischer Anwendungen. Die Frage, die sich nun stellt, lautet somit: Wie lassen sich die theoretischen Größen – die durchschnittlichen und bedingten kausalen Effekte – mittels empirisch schätzbarer Größen berechnen (*identifizieren*)?

3.4.1 Unverfälschtheit

Im Rahmen empirischer Anwendungen lassen sich bspw. Mittelwertsdifferenzen $E(Y|\,X{=}x) - E(Y|\,X{=}x')$ zwischen verschiedenen Treatment-Bedingungen x schätzen, die zunächst jedoch lediglich *Effekte auf den ersten Blick* oder *Prima-Facie-Effekte* (*PFE*; vgl. Holland, 1986) darstellen, d. h. $PFE_{xx'} \equiv E(Y|\,X{=}x) - E(Y|\,X{=}x')$. Unter Gültigkeit bestimmter Annahmen lassen sich die durchschnittlichen und bedingten kausalen Effekte über die beobachteten bedingten bzw. unbedingten Mittelwertsunterschiede, also die Prima-Facie-Effekte, identifizieren. Sind diese Annahmen in empirischen Anwendungen erfüllt, so sind die (bedingten bzw. unbedingten) Prima-Facie-Effekte *kausal unverfälscht*. Im randomisierten Experiment gilt bspw. $PFE = ACE$. Hier sind also die *unbedingten PFE* kausal unverfälscht.

Kann jedoch nicht von der kausalen Unverfälschtheit der *unbedingten PFE* ausgegangen werden – wie z. B. auch im Kontext von Schulleistungsuntersuchungen – bedarf es alternativer Datenanalysemethoden, um valide kausale Inferenzen bezüglich der durchschnittlichen Treatment-Effekte ziehen zu können. Hierbei werden nicht die unbedingten Mittelwertsunterschiede betrachtet, sondern es werden die Verfälschungen konfundierender Variablen adjustiert. Die durchschnittlichen kausalen Effekte eines Treatments werden dann basierend auf den ($\mathbf{Z}{=}\mathbf{z}$)-*bedingten PFE* (gegeben eine bestimmte Kovariatenkombination) geschätzt. Sind auch hier wiederum bestimmte Annahmen erfüllt – insbesondere solche, die die *bedingte Unverfälschtheit* der ($\mathbf{Z}{=}\mathbf{z}$)-bedingten *PFE* implizieren – erhält man eine korrekte Schätzung der entsprechenden kausalen Effekte.

Der Begriff der *Unverfälschtheit* ist sowohl für die *PFE* als auch für die jeweiligen Regressionen und deren Werte definiert (vgl. Steyer et al., 2011, Kapitel 6): Sind die bedingten Erwartungswerte $E(Y|\,X{=}x)$ für alle Werte x von X unverfälscht, so ist auch die Regression $E(Y \mid X)$ von Y auf X unverfälscht. Die Regression $E(Y \mid X)$ bezeichne

ich nachfolgend auch als *Treatment-Regression*. Da sich der *PFE* aus der Differenz dieser bedingten Erwartungswerte berechnet, d. h. $PFE_{xx'} \equiv E(Y|X{=}x) - E(Y|X{=}x')$, sind folglich auch die *PFE* unverfälscht. Entsprechendes gilt für die *bedingte Unverfälschtheit*, also die Unverfälschtheit der Regression $E(Y|X, \boldsymbol{Z})$, welche ich fortan als *Kovariaten-Treatment-Regression* bezeichne: Sind die bedingten Erwartungen $E^{X=x}(Y \mid \boldsymbol{Z})$ für alle Werte x von X unverfälscht, so ist auch die Kovariaten-Treatment-Regression $E(Y|X, \boldsymbol{Z})$ unverfälscht. Da sich der $(\boldsymbol{Z}{=}z)$-bedingte *PFE* aus der Differenz der Werte dieser Regression berechnet, d. h. $PFE_{xx';\boldsymbol{Z}=z} \equiv E(Y|X{=}x, \boldsymbol{Z}{=}z) - E(Y|X{=}x', \boldsymbol{Z}{=}z)$, sind dann auch die $(\boldsymbol{Z}{=}z)$-bedingten *PFE* unverfälscht.

Inhaltlich bedeutet die Unverfälschtheit, dass die jeweiligen *PFE* kausal interpretiert werden können. Tabelle 3.2 fasst die Definitionen der Unverfälschtheit zusammen. Hier sind in der linken Spalte die empirisch schätzbaren Größen angegeben. Diese sind genau dann unverfälscht, wenn die in der rechten Spalte aufgeführten Gleichungen gelten. Unverfälschtheit ist die schwächste Bedingung[15], welche die Berechnung kausaler Effekte ermöglicht. Sie ist jedoch keiner empirischen Überprüfung zugänglich, da die Definition der Unverfälschtheit die theoretischen Größen – die True-Outcome-Variablen τ_x sowie deren bedingte und unbedingte Erwartungswerte – beinhaltet. Daher werde ich im nachfolgenden Abschnitt ausgewählte *Kausalitätsbedingungen* einführen, die ihrerseits Unverfälschtheit implizieren und zum Teil einer empirischen Prüfung zugänglich sind.

3.4.2 Kausalitätsbedingungen

Als Kausalitätsbedingungen werden solche Bedingungen bezeichnet, die jeweils die bedingte bzw. die unbedingte Unverfälschtheit implizieren und somit eine Identifikation kausaler Effekte ermöglichen. Im Kontext der in dieser Arbeit betrachteten Schulleistungsuntersuchungen ist im Allgemeinen nicht von der *unbedingten* Unverfälschtheit der Treatment-Regression $E(Y|X)$ auszugehen. Daher betrachten wir nachfolgend ausschließlich Kausalitätsbedingungen für den $\boldsymbol{Z}$*-bedingten* Fall, d. h. die Unverfälschtheit der Kovariaten-Treatment-Regression $E(Y|X, \boldsymbol{Z})$, deren Werte und somit auch der $(\boldsymbol{Z}{=}z)$-bedingten *PFE*. Eine vollständigere Darstellung der verschiedenen Kausalitätsbedingungen findet sich bei Steyer et al. (2011, Kapitel 7 bis 9).

[15] Das Attribut *schwächste Bedingung* bedeutet, dass die Unverfälschtheit durch alle anderen Bedingungen impliziert wird, die ebenfalls die Identifikation kausaler Effekte ermöglichen.

Tabelle 3.2: Übersicht der Definitionen von Unverfälschtheit

	Empirisch schätzbare Größen	Definitionen der Unverfälschtheit
Unverfälschtheit	$E(Y \mid X{=}x)$	$E(Y \mid X{=}x) = E(\tau_x)$
	$PFE_{xx'}$	$PFE_{xx'} = E(\delta_{xx'})$
	$E(Y \mid X)$	$E(Y \mid X{=}x) = E(\tau_x)$ für jeden Wert x von X
Bedingte Unverfälschtheit[a]	$E(Y \mid X{=}x, \mathbf{Z}{=}z)$	$E(Y \mid X{=}x, \mathbf{Z}{=}z) = E(\tau_x \mid \mathbf{Z}{=}z)$
	$PFE_{xx';\mathbf{Z}=z}$	$PFE_{xx';\mathbf{Z}=z} = E(\delta_{xx'} \mid \mathbf{Z}{=}z)$
	$E^{X=x}(Y \mid \mathbf{Z})$	$E^{X=x}(Y \mid \mathbf{Z}) = E(\tau_x \mid \mathbf{Z})$
	$PFE_{xx';\mathbf{Z}}$	$PFE_{xx';\mathbf{Z}} = E(\delta_{xx'} \mid \mathbf{Z})$
	$E(Y \mid X, \mathbf{Z})$	$E^{X=x}(Y \mid \mathbf{Z}) = E(\tau_x \mid \mathbf{Z})$ für jeden Wert x von X

Anmerkungen. [a] $\mathbf{Z}$ ist die betrachtete, möglicherweise mehrdimensionale Kovariate.

Jede der im Folgenden dargestellten Kausalitätsbedingungen ist durch zwei Annahmen definiert: Neben einer jeweils spezifischen Unabhängigkeitsannahme, von denen nachfolgend vier verschiedene vorgestellt werden, wird zusätzlich angenommen, dass die Wahrscheinlichkeit, in einer der x Treatment-Bedingungen zu sein, für alle Werte z von $\mathbf{Z}$ einen Wert größer null annimmt, d. h. es gelte $P(X{=}x \mid \mathbf{Z}) > 0$ für jeden Wert x von X. Diese zweite Annahme wird auch als *Common Support*-Annahme bezeichnet (vgl. z. B. Lechner, 2000)[16].

Nach Steyer et al. ist jeder *PFE* gleich der Summe aus dem entsprechenden *ACE* und zwei verschiedenen Arten von Verzerrungen: dem *baseline bias* und dem *effect bias* (vgl. Steyer et al., 2011, Kapitel 6; Winship & Morgan, 1999). Somit sind die *PFE* dann unverfälscht, wenn (a) diese Verzerrungen beide gleich null sind (*baseline bias* = *effect bias* = 0) oder aber (b) die Verzerrungen sich gegenseitig aufheben (*baseline bias* = −*effect bias*). Diese Beziehung zwischen *PFE* und *ACE* gilt dabei für den unbedingten ebenso wie für den ($\mathbf{Z}{=}z$)-bedingten Fall. Die schwächste hinreichende Bedingung, die zur Folge hat, dass diese beiden Verzerrungen gleich null sind, ist die *$\mathbf{Z}$-bedingte regressive Unabhängigkeit der True-Outcome-Variable τ_x von der Treatment-Variable*

[16] Eine Diskussion dieser Annahme im Kontext von Schulleistungsuntersuchungen findet sich bei Fiege, 2007, Kapitel 4.

X (Abk.: $\tau_x \vdash X \mid \boldsymbol{Z},\ \forall x$)[17], d. h.:

$$E(\tau_x \mid X, \boldsymbol{Z}) = E(\tau_x \mid \boldsymbol{Z}) \qquad \text{für jeden Wert } x \text{ von } X. \tag{3.14}$$

Diese Kausalitätsbedingung wird in Abschnitt 3.5 im Rahmen der Verortung der Adjustierungsproblematik bei Vergleichsarbeiten bezüglich der Identifikation der theoretischen Parameter von zentraler Bedeutung sein.

Weitere Kausalitätsbedingungen, die jedoch jeweils die regressive Unabhängigkeit $\tau \vdash X \mid \boldsymbol{Z}$ implizieren, sollen nachfolgend lediglich exemplarisch dargestellt werden. Eine weitere Kausalitätsbedingung ist bspw. die *$\boldsymbol{Z}$-bedingte stochastische Unabhängigkeit der Treatment-Variable X und der True-Outcome-Variable* τ_x (Abk.: $X \perp\!\!\!\perp \tau_x \mid \boldsymbol{Z},\ \forall x$):

$$P(X{=}x \mid \boldsymbol{Z}, \tau_x) = P(X{=}x \mid \boldsymbol{Z}) \qquad \text{für jeden Wert } x \text{ von } X. \tag{3.15}$$

Diese Kausalitätsbedingung wird in der Literatur auch als *Strong Ignorability* (vgl. Rosenbaum & Rubin, 1983) bezeichnet. Dabei ersetzen die True-Outcome-Variablen τ_x in Gleichung 3.15 die Potential-Outcome-Variablen Y_x in der Rubinschen Terminologie (vgl. Abschnitt 3.2). Der Nachteil dieser Kausalitätsbedingung ist jedoch, dass diese – ebenso wie die $\boldsymbol{Z}$-bedingte regressive Unabhängigkeit der True-Outcome-Variable τ_x von X – einer empirischen Prüfung nicht zugänglich ist. Daher stelle ich nun exemplarisch zwei weitere Bedingungen vor, die einer empirischen Prüfung zugänglich sind bzw. im Rahmen des Designs einer Untersuchung hergestellt werden können:

(1) Die erste Bedingung ist die *$\boldsymbol{Z}$-bedingte stochastische Unabhängigkeit der Treatment-Variable X und der umfassenden Kovariate* C_X (Abk.: $X \perp\!\!\!\perp C_X \mid \boldsymbol{Z}$). Diese Bedingung ist genau dann erfüllt, wenn gilt:

$$P(X{=}x \mid \boldsymbol{Z}, C_X) = P(X{=}x \mid \boldsymbol{Z}) \qquad \text{für jeden Wert } x \text{ von } X. \tag{3.16}$$

Falls $C_X = U$ ist diese Bedingung bspw. dann erfüllt, wenn die individuellen Treatment-Wahrscheinlichkeiten bei gegebenem Wert $\boldsymbol{z}$ einer Kovariaten $\boldsymbol{Z}$ für alle Beobachtungseinheiten gleich sind[18]. Diese Bedingung ist im Rahmen von experimentellen Untersuchungsdesigns mittels (bedingter) Randomisierung herstellbar. In Beobachtungsstudien hingegen kann diese Bedingung durch die Aus-

[17] Das Symbol $\forall x$ bedeutet *für alle Werte* x.
[18] In diesem Fall gilt: $P(X{=}x \mid \boldsymbol{Z}, U) = P(X{=}x \mid \boldsymbol{Z})$ für jeden Wert x von X.

wahl der Kovariaten $Z_1, \ldots, Z_Q$ hergestellt werden, die darauf abzielt, dass die bedingte Unabhängigkeit von X und C_X gegeben $\boldsymbol{Z} = (Z_1, \ldots, Z_Q)$ zutrifft.

(2) Die zweite Bedingung ist die *$\boldsymbol{Z}$-bedingte regressive Unabhängigkeit der Outcome-Variable Y von C_X gegeben X* (Abk.: $Y \vdash C_X \mid X, \boldsymbol{Z}$), d. h.:

$$E(Y \mid X, C_X) = E(Y \mid X, \boldsymbol{Z}) . \tag{3.17}$$

Für den Fall, dass $C_X = U$, bedeutet dies inhaltlich: Bei gegebenem Wert $\boldsymbol{z}$ der Kovariaten $\boldsymbol{Z}$ gibt es keine interindividuellen Unterschiede zwischen den Beobachtungseinheiten bezogen auf die bedingten Erwartungswerte der Outcome-Variablen Y in jeder der Treatment-Stufen[19]. Alle Personen mit einer bestimmten Kovariatenkonstellation $\boldsymbol{z} = (z_1, \ldots, z_Q)$ in einer konkreten Behandlungsbedingung x haben also den gleichen Erwartungswert bezüglich der Variablen Y.

Diese beiden Bedingungen – (1) $X \perp\!\!\!\perp C_X \mid \boldsymbol{Z}$ und (2) $Y \vdash C_X \mid X, \boldsymbol{Z}$ – zeichnen sich dadurch aus, dass sie in empirischen Anwendungen getestet werden können[20], zumindest im Sinne einer möglichen Falsifikation dieser Annahmen (vgl. z. B. Mayer et al., 2011). Die Unverfälschtheit (vgl. Tabelle 3.2) selbst jedoch kann in empirischen Anwendungen weder falsifiziert noch verifiziert werden. Wichtig ist weiterhin, dass sowohl Bedingungen (1) als auch (2) jeweils die bedingte regressive Unabhängigkeit $\tau_x \vdash X \mid \boldsymbol{Z}$, $\forall x$ der True-Outcome-Variablen τ_x von der Treatment-Variablen X gegeben $\boldsymbol{Z}$ – und damit auch die Unverfälschtheit der Kovariaten-Treatment-Regression $E(Y \mid X, \boldsymbol{Z})$ – implizieren. Tabelle 3.3 fasst die im Rahmen dieser Arbeit dargestellten vier Kausalitätsbedingungen zusammen, d. h. deren Definition, Beschreibung sowie die im weiteren Verlauf dieser Arbeit verwendeten Kurzformen.

3.5 Kausale Effekte und faire Vergleiche in Vergleichsarbeiten

Will man nun faire Vergleiche aus Vergleichsarbeiten als kausale Effekte des Unterrichts auf die Testleistung der Schüler interpretieren, so müssen sich diese fairen Ver-

[19] Dann gilt also: $E^{X=x}(Y \mid \boldsymbol{Z}, U) = E^{X=x}(Y \mid \boldsymbol{Z})$ für jeden Wert x von X.

[20] Weitere Kausalitätsbedingungen, die ebenfalls einer empirischen Prüfung zugänglich sind, finden sich bei Steyer et al., 2011, Kapitel 7 bis 9.

Tabelle 3.3: Vier ausgewählte Kausalitätsbedingungen, die jeweils die $\boldsymbol{Z}$-bedingte Unverfälschtheit implizieren

Definition[a]	Bedeutung	Kurzform
$E(\tau_x \mid X, \boldsymbol{Z}) = E(\tau_x \mid \boldsymbol{Z})$ für jeden Wert x von X	$\boldsymbol{Z}$-bedingte regressive Unabhängigkeit der True-Outcome-Variable τ_x von X	$\tau_x \vdash X \mid \boldsymbol{Z}, \forall x$
$P(X{=}x \mid \boldsymbol{Z}, \tau_x) = P(X{=}x \mid \boldsymbol{Z})$ für jeden Wert x von X	$\boldsymbol{Z}$-bedingte stochastische Unabhängigkeit der Treatment-Variable X und τ_x	$X \perp\!\!\!\perp \tau_x \mid \boldsymbol{Z}, \forall x$
$P(X{=}x \mid \boldsymbol{Z}, C_X) = P(X{=}x \mid \boldsymbol{Z})$ für jeden Wert x von X	$\boldsymbol{Z}$-bedingte stochastische Unabhängigkeit der Treatment-Variable X und C_X	$X \perp\!\!\!\perp C_X \mid \boldsymbol{Z}$
$E(Y \mid X, C_X) = E(Y \mid X, \boldsymbol{Z})$	$\boldsymbol{Z}$-bedingte regressive Unabhängigkeit der Outcome-Variable Y von C_X gegeben X	$Y \vdash C_X \mid X, \boldsymbol{Z}$

Anmerkungen. [a]Jede der vier Definitionen enthält zusätzlich die *Common Support*-Annahme, d. h., es gelte jeweils $P(X{=}x \mid \boldsymbol{Z}) > 0$ für jeden Wert x von X.

gleiche kausaltheoretisch verorten lassen. Dies ist dann der Fall, wenn die empirisch schätzbaren Parameter, die zur Berechnung fairer Vergleiche verwendet werden, mit den theoretischen Parametern der allgemeinen stochastischen Theorie kausaler Effekte übereinstimmen und die dafür ggf. erforderlichen Annahmen expliziert werden können. Diese kausaltheoretische Einordnung – und damit die Frage nach der *Fairness fairer Vergleiche* – ist Inhalt des folgenden Abschnitts, der im Wesentlichen auf Fiege (2007) basiert. In Fiege (2007) wird ausführlich die formale Repräsentation des Adjustierungsvorgehens im Projekt *Kompetenztest.de* und PISA-E 2000 zur Berechnung fairer Vergleiche sowie deren kausaltheoretische Verortung dargestellt. Nachfolgend werde ich daraus die für die vorliegende Arbeit zentralen Ergebnisse herausgreifen, zusammenfassen sowie ergänzen.

3.5.1 Der intendierte kausale Effekt: Definition und Identifikation

Wie bereits in den Abschnitten 3.3.2 und 3.3.3 dargestellt, wird der Lehrer einer Klasse in aller Regel die Optimierung der Unterrichtseffekte bezüglich der jeweils *eigenen* Schülerschaft intendieren und nicht bezüglich Schüler im Allgemeinen, die zu großen Teilen gar nicht als Schüler seiner Klasse in Frage kommen. Daher zielen wir im Kontext von Vergleichsarbeiten auf die Schätzung des ($X{=}x$)-bedingten kausalen Effekts

$CCE_{x;\,X=x}$ von x, d. h. des *ACE on the treated*, wobei jeder Wert x von X den Unterricht einer Klasse repräsentiert. Anders als im klassischen Design eines randomisierten Experiments wird im Rahmen des vorliegenden Beobachtungsstudiendesigns jedoch keine Kontrollbedingung im üblichen Sinne, d. h. *ohne* Treatment, betrachtet. Dies trifft insbesondere für Vergleichsarbeiten zu, die im Fokus dieser Arbeit stehen: Jede Beobachtungseinheit (jeder Schüler) wird dem Unterricht einer Klasse und damit einer der J Treatment-Stufen $j = 1, \ldots, J$ zugewiesen. Keiner der Schüler erhält *kein* Treatment (keinen Unterricht). Da bei Vergleichsarbeiten zudem nicht nur zwei, sondern insgesamt J Treatment-Bedingungen verglichen werden, steht nachfolgend die Effektparametrisierung des *ACE on the treated* im Fokus der Betrachtung (vgl. Abschnitt 3.3.3). Ziel ist somit die Identifikation und Schätzung des folgenden bedingten Effekts:

$$
\begin{aligned}
CCE_{x;\,X=x} &\equiv E(\delta_x \mid X{=}x) && (3.18)\\
&= E\Big(\tau_x - J^{-1} \sum_{x'=1}^{J} \tau_{x'} \mid X{=}x\Big) && \text{[Gleichung 3.9]} \quad (3.19)\\
&= E(\tau_x \mid X{=}x) - J^{-1} \sum_{x'=1}^{J} E(\tau_{x'} \mid X{=}x)\,, && (3.20)
\end{aligned}
$$

wobei $CCE_{x;\,X=x}$ der bedingte kausale Effekt von x verglichen mit dem Durchschnitt aller Treatment-Bedingungen gegeben Treatment-Bedingung $X{=}x$ ist. Im Kontext von Vergleichsarbeiten interessiert uns somit der $CCE_{x;\,X=x}$, d. h. der $(X{=}x)$-bedingte Effekt des Unterrichts einer Klasse x im Vergleich zum durchschnittlichen Unterricht in allen Klassen. Doch wie kann diese theoretische Größe in empirischen Anwendungen identifiziert werden?

Zum Zwecke der Identifikation des $(X{=}x)$-bedingten kausalen Effekts $CCE_{x;\,X=x}$ betrachten wir nachfolgend die Differenz in Gleichung 3.20 im Detail. Für den Minuenden aus Gleichung 3.20 gilt ohne weitere Annahmen:

$$
\begin{aligned}
E(\tau_x \mid X{=}x) &= E\big[\, E^{X=x}(Y \mid C_X) \mid X{=}x \,\big] && \text{[Gleichung 3.1]} \quad (3.21)\\
&= E^{X=x}\big[\, E^{X=x}(Y \mid C_X) \,\big] && \text{[RR (iv) für Regressionen}^{21}\text{]}\\
&= E(Y \mid X{=}x)\,,
\end{aligned}
$$

d. h., der Erwartungswert $E(Y \mid X{=}x)$ der Outcome-Variable Y in einer konkreten Klas-

[21] Rechenregel (iv) für Regressionen: $E[E(Y \mid X)] = E(Y)$ (Steyer, 2003, S. 85).

se x ist gleich dem Erwartungswert $E(\tau_x \mid X{=}x)$ der True-Outcome-Variable τ_x gegeben Klasse x (vgl. Fiege, 2007, Kapitel 4). Somit ist der Minuend mittels des empirisch schätzbaren Parameters $E(Y \mid X{=}x)$ identifiziert, der wiederum in empirischen Anwendungen über den beobachteten Mittelwert der Testleistungen einer Klasse x geschätzt werden kann.

Zur Identifikation des Subtrahenden aus Gleichung 3.20 bedarf es hingegen weiterer Annahmen, wobei eine der Kausalitätsbedingungen für diesen Fall hinreichend ist (vgl. Fiege, 2007, Kapitel 4): Kann von der $\boldsymbol{Z}$-bedingten regressiven Unabhängigkeit der True-Outcome-Variable τ_x von X ($\tau_x \vdash X \mid \boldsymbol{Z}, \forall x$) ausgegangen werden – d. h. es gelte $E(\tau_x \mid X, \boldsymbol{Z}) = E(\tau_x \mid \boldsymbol{Z})$ für jeden Wert x von X –, so folgt daraus die bedingte Unverfälschtheit der Kovariaten-Treatment-Regression $E(Y \mid X, \boldsymbol{Z})$. Dies wiederum impliziert:

$$\begin{aligned} E(\tau_{x'} \mid X) &= E\big[\, E(\tau_{x'} \mid X, \boldsymbol{Z}) \mid X \,\big] && \text{[RR (vi) für Regressionen}^{22}\text{]} \\ &= E\big[\, E(\tau_{x'} \mid \boldsymbol{Z}) \mid X \,\big] && [\tau_x \vdash X \mid \boldsymbol{Z}, \forall x] \\ &= E\big[\, E^{X=x'}(Y \mid \boldsymbol{Z}) \mid X \,\big]. \end{aligned} \tag{3.22}$$

Falls $\tau_x \vdash X \mid \boldsymbol{Z}, \forall x$ zutrifft, ist der $(X{=}x)$-bedingte kausale Effekt $CCE_{x;\,X=x}$ von x, d. h. der durchschnittliche kausale Effekt von Treatment x verglichen mit dem Durchschnitt aller Treatment-Bedingungen gegeben Treatment-Bedingung $X{=}x$, somit wie folgt identifiziert:

$$CCE_{x;\,X=x} = E(\tau_x \mid X{=}x) - J^{-1} \sum_{x'=1}^{J} E(\tau_{x'} \mid X{=}x) \tag{3.23}$$

$$= E(Y \mid X{=}x) - J^{-1} \sum_{x'=1}^{J} E\big[\, E^{X=x'}(Y \mid \boldsymbol{Z}) \mid X{=}x \,\big] \tag{3.24}$$

$$= E(Y \mid X{=}x) - Ref_{causal}\,. \tag{3.25}$$

Der Subtrahend in Gleichung 3.24 ist der adjustierte, kausal interpretierbare Referenzwert, der nachfolgend mit Ref_{causal} (vgl. Gleichung 3.25) bezeichnet wird, d. h. es gelte: $Ref_{causal} \equiv J^{-1} \sum_{x'=1}^{J} E\big[\, E^{X=x'}(Y \mid \boldsymbol{Z}) \mid X{=}x \,\big]$.

[22] Rechenregel (vi) für Regressionen: $E[E(Y \mid X) \mid f(X)] = E[Y \mid f(x)]$ (Steyer, 2003, S. 85).

Purged Conditional Expectations. Die Gleichungen 3.22 sowie 3.24 zeigen zudem die Rationale, auf der statistische Adjustierungsverfahren basieren. Ganz allgemein formuliert haben statistische Adjustierungsverfahren das Ziel, die durch konfundierende Variablen resultierenden Verfälschungen zu bereinigen. Nachfolgend sei W eine zunächst beliebige Zufallsvariable, um deren potenziellen Einfluss der Erwartungswert einer Outcome-Variable Y in einer Treatment-Bedingung $X{=}x$ bereinigt werden soll. Dann ist $\overline{E}^{W}(Y|\, X{=}x)$ der W-bereinigte $(X{=}x)$-bedingte Erwartungswert von Y (*W-purged $(X{=}x)$-conditional expectation of Y*; vgl. Steyer, Nagel et al., in Druck), der wie folgt definiert ist:

$$\overline{E}^{W}(Y|\, X{=}x) \;\equiv\; E[\, E^{X=x}(Y|\, W\,)\,]\,. \tag{3.26}$$

Werden also die bedingten Erwartungswerte $E^{X=x}(Y \mid W{=}w)$ über die (unbedingte) Verteilung von W aggregiert, spricht man von *W-bereinigten $(X{=}x)$-bedingten Erwartungswerten* (vgl. Steyer, Nagel et al., in Druck) und deren Differenzen, die jedoch nicht ohne weitere Annahmen kausal interpretierbar sind. Sind die W-bereinigten $(X{=}x)$-bedingten Erwartungswerte zudem kausal unverfälscht (vgl. auch Abschnitt 3.4.1), so lässt sich mittels der auf diese Weise bereinigten Erwartungswerte gleichfalls der entsprechende kausale Effekt identifizieren. Ein Spezialfall davon liegt bspw. vor, wenn W *alle* konfundierenden Kovariaten umfasst, d. h. alle der Treatment-Variable vor- oder gleichgeordnete Variablen, welche die Outcome-Variable Y über das Treatment hinaus beeinflussen können. In diesem Fall ist W gleichzusetzen mit der umfassenden Kovariate C_X und es gilt: $W = C_X$. Daraus folgt $\overline{E}^{W}(Y|\, X{=}x) \;\equiv\; E[\, E^{X=x}(Y|\, W\,)\,] = E[\, E^{X=x}(Y|\, C_X\,)\,] = E(\tau_x)$, d. h. der W-bereinigte $(X{=}x)$-bedingte Erwartungswert von Y ist kausal unverfälscht. Sind die Kausalitätsbedingungen jedoch nicht erfüllt und kann nicht von der bedingten Unverfälschtheit ausgegangen werden, so liefern auch statistische Adjustierungsverfahren keine kausalen Effektschätzungen.

3.5.2 Der adjustierte Effekt $E(\delta_{adj} \mid X{=}x)$ in Vergleichsarbeiten und seine kausaltheoretische Verortung: Kausal oder nicht kausal, das ist hier die Frage

Fiege (2007) stellt weiterhin das tatsächlich praktizierte Vorgehen im Rahmen von Vergleichsarbeiten bzw. allgemein Schulleistungsuntersuchungen dar. Zu diesem Zweck werden zwei Beispiele herangezogen: Das Adjustierungsverfahren im Projekt *Kompe-*

tenztest.de sowie das Adjustierungsverfahren in der nationalen Erweiterung von PISA 2000 (PISA-E 2000). Es wird gezeigt, dass zur Berechnung eines klassenspezifischen Effektmaßes – dem potenziell *fairen Vergleich* – bei der Adjustierungsstrategie vom Projekt *Kompetenztest.de* eine saturierte Parametrisierung verwendet wird. Im Gegensatz dazu wurde im Kontext von PISA-E 2000 eine lineare Parametrisierung zur Berechnung verwendet. Unabhängig von der konkreten Parametrisierung (saturiert vs. linear), ist der adjustierte Effekt $E(\delta_{adj} \mid X{=}x)$ wie folgt definiert:

$$\begin{aligned} E(\delta_{adj} \mid X{=}x) &\equiv E(Y \mid X{=}x) - E\big[\, E(Y \mid \mathbf{Z}) \mid X{=}x \,\big] \\ &= E(Y \mid X{=}x) - Ref_{adj} \,, \end{aligned} \tag{3.27}$$

wobei $\mathbf{Z}$ der Vektor einer Q-dimensionalen Kovariaten $\mathbf{Z} = (Z_1, \ldots, Z_Q)$ ist. Weiterhin sei $Ref_{adj} \equiv E\big[\, E(Y \mid \mathbf{Z}) \mid X{=}x \,\big]$ der adjustierte Referenzwert. Im Kontext von *Value-Added Modellen* (VAM) wird dieser adjustierte Effekt $E(\delta_{adj} \mid X{=}x)$ häufig auf Ebene der Schulen berechnet (vgl. Kapitel 4). Im Projekt *Kompetenztest.de* wird $E(\delta_{adj} \mid X{=}x)$ für den Unterricht einzelner Klassen x, d. h. auf Klassenebene, berechnet. Die Schätzung dieser klassenspezifischen Effekte erfolgt schrittweise. Die einzelnen Analyseschritte werden ausführlich in Kapitel 6 (Abschnitt 6.3) dargestellt.

Das bedeutet jedoch nicht, dass $E(\delta_{adj} \mid X{=}x)$ ohne weitere Annahmen kausal interpretiert werden kann. Zwar ist der Minuend aus Gleichung 3.27 ein Schätzer für den Minuenden aus Gleichung 3.20, d. h. es gilt: $E(Y \mid X{=}x) = E(\tau_x \mid X{=}x)$. Dennoch gilt nicht ohne Weiteres $E(\delta_{adj} \mid X{=}x) = CCE_{x;\, X=x}$. Mit anderen Worten: Der adjustierte klassenspezifische Effekt $E(\delta_{adj} \mid X{=}x)$ ist nicht ohne weitere Annahmen identisch mit dem *ACE on the treated*. Das ist nur dann der Fall, wenn der Subtrahend aus Gleichung 3.27 mit dem Subtrahenden aus Gleichung 3.24 übereinstimmt, wenn also $Ref_{adj} = Ref_{causal}$.

Zunächst gilt jedoch:

$$\begin{aligned} Ref_{adj} &\neq Ref_{causal} \\ E\big[\, E(Y \mid \mathbf{Z}) \mid X{=}x \,\big] &\neq J^{-1} \sum_{x'=1}^{J} E\big[\, E^{X=x'}(Y \mid \mathbf{Z}) \mid X{=}x \,\big] \,. \end{aligned} \tag{3.28}$$

Damit die Äquivalenzbeziehung $Ref_{adj} = Ref_{causal}$ zutrifft – und folglich auch das adjustierte Effektmaß $E(\delta_{adj} \mid X{=}x)$ kausal interpretierbar ist –, müssen weitere, sehr starke Annahmen erfüllt sein. So gilt für den im Kontext von Vergleichsarbeiten betrachteten

adjustierten Referenz- bzw. Vergleichswert Ref_{adj}[23]:

$$\begin{aligned} Ref_{adj} &= E\big[\, E(Y \mid \boldsymbol{Z}) \mid X{=}x \,\big] && (3.29)\\ &= \sum_{z} E(Y \mid \boldsymbol{Z}{=}z) \cdot P(\boldsymbol{Z}{=}z \mid X{=}x) && \text{[RR (iv)}^{24}\text{]}\\ &= \sum_{z} \sum_{x'=1}^{J} E^{X=x'}(Y \mid \boldsymbol{Z}{=}z) \cdot P(X{=}x' \mid \boldsymbol{Z}{=}z) \cdot P(\boldsymbol{Z}{=}z \mid X{=}x)\ . \end{aligned}$$

Für den kausaltheoretischen Referenzwert Ref_{causal} hingegen gilt:

$$\begin{aligned} Ref_{causal} &= J^{-1} \sum_{x'=1}^{J} E\big[\, E^{X=x'}(Y \mid \boldsymbol{Z}) \mid X{=}x \,\big] && (3.30)\\ &= J^{-1} \sum_{x'=1}^{J} \sum_{z} E^{X=x'}(Y \mid \boldsymbol{Z}{=}z) \cdot P(\boldsymbol{Z}{=}z \mid X{=}x) && \text{[RR für Summen}^{25}\text{]}\\ &= \sum_{z} \sum_{x'=1}^{J} E^{X=x'}(Y \mid \boldsymbol{Z}{=}z) \cdot J^{-1} \cdot P(\boldsymbol{Z}{=}z \mid X{=}x)\ . \end{aligned}$$

Gleichung 3.29 und 3.30 sind äquivalent, wenn gilt:

$$P(X{=}x' \mid \boldsymbol{Z}{=}z) = J^{-1}\ , \qquad (3.31)$$

d. h., falls die ($\boldsymbol{Z}{=}z$)-bedingte Treatment-Wahrscheinlichkeit für alle Werte z von $\boldsymbol{Z}$ und x von X identisch sind. Dies setzt jedoch voraus, dass „ … nicht nur die Anzahl der Schüler, sondern auch die Verteilungen der Kovariaten in allen […] Klassen identisch sind, wovon im Rahmen der vorliegenden Anwendungen [Schulleistungsuntersuchungen] jedoch nicht ausgegangen werden kann“ (Fiege, 2007, S. 36).

Somit ist der auf diese Weise berechnete Referenzwert $E\big[\, E(Y \mid \boldsymbol{Z}) \mid X{=}x \,\big]$ zunächst lediglich ein *deskriptives Maß*, d. h. der für eine Klasse bzw. Schule zu erwartende Wert adjustiert für die konkreten Variablen $\boldsymbol{Z}$ im Adjustierungsmodell. Werden in ei-

[23] In der vorliegenden Arbeit verwende ich das Summenzeichen. Diese spezielle Integration ist immer dann zutreffend, wenn es sich im diskrete Zufallsvariablen handelt. Im Falle kontinuierlicher Zufallsvariablen muss entsprechend integriert werden.

[24] Rechenregel (iv) für bedingte Erwartungswerte: $E(Y|X{=}x) = \sum_z E(Y|X{=}x, Z{=}z) \cdot P(Z{=}z|X{=}x)$. Diese Regel gilt immer dann, falls Z diskret ist *und* falls gilt: $P(X{=}x, Z{=}z) > 0,\ \ \forall x, z \in \Omega_X \times \Omega_Z$ (Steyer, 2003, S. 81).

[25] Rechenregel für Doppelsummen und Rechenregel für Summen mit konstantem Faktor: $\sum_i (c \cdot a_i) = c \cdot \sum_i a_i$.

nem nächsten Schritt zusätzliche Variablen $\boldsymbol{Z}^+$ im Adjustierungsmodell berücksichtigt, so kann sich ein anderer Referenzwert ergeben. Der nun resultierende Vergleichswert $E[\,E(Y\,|\,\boldsymbol{Z},\boldsymbol{Z}^+)\,|\,X{=}x\,]$ ist der für eine Klasse bzw. Schule zu erwartende Wert adjustiert für $\boldsymbol{Z}$ *und* $\boldsymbol{Z}^+$ im Adjustierungsmodell. Dieser kann sich somit von dem zuerst berechneten adjustierten Vergleichswert $E[\,E(Y\,|\,\boldsymbol{Z})\,|\,X{=}x\,]$ unterscheiden. Zudem werden sich beide Werte von dem kausalen Vergleichs- bzw. Referenzwert unterscheiden, sofern nicht die Annahme der bedingten Unverfälschtheit erfüllt ist *und* Gleichung 3.31 gilt.

Zusammenfassend lässt sich feststellen, dass eine kausaltheoretische Verortung des adjustierten Effektmaßes zwar möglich ist. In Anwendungen ist die Schätzung fairer Vergleiche im Sinne kausaler Unterrichtseffekte jedoch nicht möglich, da die dazu erforderlichen Annahmen im Kontext von Vergleichsarbeiten nicht haltbar sind. Mittels der praktizierten Adjustierungen lässt sich bestenfalls eine Näherung an die intendierten kausalen Effekte erreichen. Daher verwende ich nachfolgend den Begriff *fairere Vergleiche* (vgl. Fiege et al., 2011; Nachtigall et al., 2009).

Exkurs: Eine alternative Definition des *ACE on the treated*. Zwecks Einordnung der Adjustierungsproblematik bei Vergleichsarbeiten im Rahmen der Kausalitätstheorie verwendet Fiege (2007) außerdem eine zweite, alternative Effektdefinition. Wie bereits ausführlich dargestellt wird bei der ersten Definition (Effektparametrisierung des *ACE on the treated*) jeweils ein Treatment, also der Unterricht in einer Klasse $X{=}x$, mit dem Durchschnitt aller Treatment-Bedingungen verglichen, wobei jede der mittels der Vergleichsarbeiten untersuchten Klassen eine Treatment-Stufe repräsentiert. In der zweiten, alternativen Definition hingegen wird eine vereinfachte Sichtweise eingenommen, bei der nur jeweils zwei Treatment-Gruppen $X{=}x$ und $X{=}x'$ betrachtet werden (binäre Treatment-Definition). Dabei sei $X{=}x$ der Unterricht in einer betrachteten Klasse und $X{=}x'$ der Unterricht in *allen anderen* Klassen $X \neq x$. Hier findet die Differenzparametrisierung des *ACE on the treated* Anwendung (vgl. Gleichung 3.8). Jedoch müssen auch im Kontext der kausaltheoretischen Verortung des auf diese Weise definierten Effektmaßes zusätzliche Annahmen getroffen werden, die über die Annahme der (bedingten) Unverfälschtheit hinausgehen (vgl. Fiege, 2007, Kapitel 4). Zudem ist es unplausibel, den Unterricht in den verschiedenen Klassen als *einen gemeinsamen* kausalen Agens aufzufassen. Eine binäre Treatment-Definition im Kontext des vorliegenden Anwendungsfalls ist somit nicht haltbar.

3.5.3 Eigenschaften der adjustierten Effektfunktion $E(\delta_{adj} \mid X)$

Im Folgenden betrachten wir die X-bedingte adjustierte Effektfunktion $E(\delta_{adj}|X)$, deren Werte die adjustierten klassenspezifischen Effekte $E(\delta_{adj} \mid X{=}x)$ sind. Eine Eigenschaft der adjustierten Effektfunktion $E(\delta_{adj} \mid X)$ ist, dass deren theoretischer Mittelwert über alle Klassen $X{=}x$ den Wert null annimmt:

$$
\begin{aligned}
E[\, E(\delta_{adj} \mid X)\,] &= E\Big[\; E(Y|X) - E[E(Y \mid \boldsymbol{Z}) \mid X]\,\Big] && \text{[Gleichung 3.27]} \\
&= E[\, E(Y - E(Y \mid \boldsymbol{Z}) \mid X)\,] && \text{[RR (iii) für Erwartungswerte}^{26}\text{]} \\
&= E[\, E(\varepsilon_{E(Y|\boldsymbol{Z})} \mid X)\,] && \text{[Definition Residuum}^{27}\text{]} \\
&= E[\, \varepsilon_{E(Y|\boldsymbol{Z})}\,] && \text{[RR (iv) für Regressionen}^{28}\text{]} \\
&= 0\,. && \text{[Eigenschaft (ii) des Residuums}^{29}\text{]}
\end{aligned}
\tag{3.32}
$$

Gilt dies auch für die entsprechenden kausalen Effekte? Und trifft diese Eigenschaft unabhängig davon zu, ob die Differenz- oder die Effektparametrisierung verwendet wird?

Betrachten wir zunächst die Differenzparametrisierung (vgl. Abschnitt 3.3.2 und Tabelle 3.1): Für den Erwartungswert der X-bedingten kausalen Effektfunktion $E(\delta_{xx'}|X)$, deren Werte die $(X{=}x^*)$-bedingten kausalen Effekte von x vs. x' sind, gilt zunächst nur:

$$
E[\, E(\delta_{xx'} \mid X)\,] = E(\delta_{xx'})\,. \tag{3.33}
$$

Das bedeutet, dass der Erwartungswert der X-bedingten kausalen Effektfunktion gleich dem durchschnittlichen kausalen Effekt $ACE_{xx'}$ von x vs. x' ist (vgl. Gleichung 3.5). Dieser kann zwar in empirischen Anwendungen den Wert null annehmen, er ist jedoch nicht *per definitionem* null.

Auch für den Fall der Effektparametrisierung kausaler Effekte (vgl. Abschnitt 3.3.3 und Tabelle 3.1) lässt sich zeigen, dass der theoretische Mittelwert über die bedingten Effekte nicht per definitionem den Wert null annimmt. So gilt für den Erwartungswert der X-bedingten kausalen Effektfunktion $E(\delta_x \mid X)$, deren Werte die $(X{=}x)$-bedingten

[26] Rechenregel (iii) für Erwartungswerte: $E(\alpha \cdot Y_1 + \beta \cdot Y_2) = \alpha \cdot E(Y_1) + \beta \cdot E(Y_2)$ (Steyer, 2003, S. 61).
[27] Definition des Residuums ε bezüglich der Regression $E(Y \mid X)$: $\varepsilon = Y - E(Y \mid X)$ (Steyer, 2003, S. 86).
[28] Rechenregel (iv) für Regressionen: $E[E(Y \mid X)] = E(Y)$ (Steyer, 2003, S. 85).
[29] Eigenschaft (ii) des Residuums: $E(\varepsilon) = 0$ (Steyer, 2003, S. 89).

kausalen Effekte $CCE_{x;\,X=x}$ von x sind, zunächst folgende Gleichung:

$$
\begin{aligned}
E(CCE_{x;\,X}) &= E[\,E(\delta_x \mid X)\,] && (3.34)\\
&= E[\,E(\tau_x \mid X) - J^{-1}\sum_{x'=1}^{J} E(\tau_{x'} \mid X)\,] && \text{[Gleichung 3.20]}\\
&= E[\,E(\tau_x \mid X)\,] - E[\,J^{-1}\sum_{x'=1}^{J} E(\tau_{x'} \mid X)\,] && \text{[RR (iii) für Erwartungswerte}^{30}\text{]}\\
&= E[\,E(\tau_x \mid X)\,] - J^{-1}\sum_{x'=1}^{J} E[\,E(\tau_{x'} \mid X)\,] && \text{[RR (iii) für Erwartungswerte}^{30}\text{]}\\
&= E(\tau_x) - J^{-1}\sum_{x'=1}^{J} E(\tau_{x'}) && \text{[RR (iv) für Regressionen}^{31}\text{]}\\
&= E(\,\tau_x - J^{-1}\sum_{x'=1}^{J} \tau_{x'}\,) && \text{[RR (iii) für Erwartungswerte}^{30}\text{]}\\
&= E(\delta_x) && \text{[Gleichung 3.9]}\\
&= ACE_x\,. && \text{[Gleichung 3.10]}
\end{aligned}
$$

Der Erwartungswert der X-bedingten kausalen Effektfunktion $E(\delta_x \mid X)$ ist somit gleich dem durchschnittlichen kausalen Effekt ACE_x von x. Man beachte an dieser Stelle, dass die Referenzbedingung – in diesem Fall die Treatment-Stufe $X{=}x$ – hierbei jedoch stets unverändert bleibt.

Dies soll an einem fiktiven Beispiel verdeutlicht werden, welches in Abbildung 3.3 dargestellt ist. Hierbei betrachten wir nur zwei Treamtent-Stufen $X = a$ und $X = b$, die jeweils den Unterricht in einer von zwei Klassen repräsentieren. Die linke Seite in Abbildung 3.3 zeigt die vier möglichen ($X{=}x$)-bedingten Erwartungswerte $E(\tau_x \mid X{=}x)$ der beiden True-Outcome-Variablen τ_a und τ_b, wobei die zwei Spalten die jeweilige Bedingung (d. h. $X = a$ bzw. $X = b$) repräsentieren. ($X{=}x$)-bedingte Erwartungswerte, die eine positive Abweichung vom zugehörigen kausalen Referenzwert Ref_{causal} aufweisen, sind grün markiert. Werte mit negativer Abweichung sind in Rot dargestellt. Außerdem ist der jeweilige bedingte kausale Referenzwert Ref_{causal} in Grau eingetragen. Im vorliegenden Minimalbeispiel gibt es zwei bedingte kausale Referenzwerte: $Ref_{causal;\,a}$ und $Ref_{causal;\,b}$. Auf der rechten Seite von Abbildung 3.3 sind nun zusätzlich

[30] Rechenregel (iii) für Erwartungswerte: $E(\alpha \cdot Y_1 + \beta \cdot Y_2) = \alpha \cdot E(Y_1) + \beta \cdot E(Y_2)$ (Steyer, 2003, S. 61).
[31] Rechenregel (iv) für Regressionen: $E[E(Y \mid X)] = E(Y)$ (Steyer, 2003, S. 85).

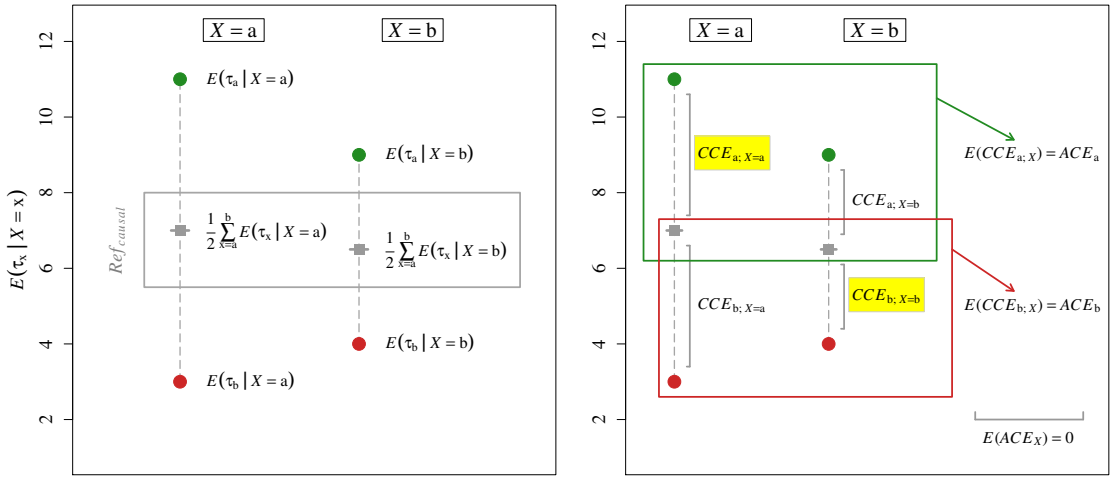

Abbildung 3.3: Minimalbeispiel mit lediglich zwei Treatment-Stufen $X = a$ und $X = b$. Links: $(X{=}x)$-bedingte Erwartungswerte $E(\tau_x|X{=}x)$ der True-Outcome-Variablen τ_x und der jeweilige kausale Referenzwert Ref_{causal}. Rechts: $(X{=}x)$-bedingte kausale Effekte $CCE_{x;\,X=x}$ von x und deren Erwartungswerte. Es gelte jeweils $P(X{=}x) = J^{-1}$ für jeden Wert x von X.

die Abweichungen der $(X{=}x)$-bedingten Erwartungswerte $E(\tau_x|X{=}x)$ von dem jeweiligen bedingten kausalen Referenzwert gekennzeichnet. Diese Abweichungen sind die $(X{=}x)$-bedingten kausalen Effekte $CCE_{x;\,X=x}$ von x. Gelb markiert sind die zwei Effekte, die wir im Kontext von Vergleichsarbeiten schätzen wollen[32] und an den Lehrer in Klasse a bzw. den Lehrer in Klasse b zurückmelden: der $CCE_{a;\,X=a}$ und der $CCE_{b;\,X=b}$.

Mittelt man über die $(X{=}x)$-bedingten kausalen Effekte $CCE_{x;\,X=x}$ von x *zwischen* den Treatment-Bedingungen $X = a$ und $X = b$, so erhält man den durchschnittlichen kausalen Effekt ACE_x von x (vgl. Gleichung 3.34). In unserem Beispiel können wir wiederum exakt zwei solcher theoretische Mittelwerte betrachten:

(1) den ACE_a, d. h. den Erwartungswert $E(CCE_{a;\,X})$ der beiden $(X{=}x)$-bedingten kausalen Effekte $CCE_{a;\,X=x}$ vom Unterricht in Klasse a im grünen Rechteck *und*

(2) den ACE_b, d. h. Erwartungswert $E(CCE_{b;\,X})$ der $(X{=}x)$-bedingten kausalen Effekte $CCE_{b;\,X=x}$ vom Unterricht in Klasse b im roten Rechteck.

Im vorliegenden Beispiel sind beide kausalen Effekte $ACE_x \neq 0$: Der ACE_a nimmt einen postiven Wert an, wohingegen der ACE_b ein negatives Vorzeichen erhält.

[32] Dies ist dann der Fall, wenn die in Abschnitt 3.5.2 dargestellten Annahmen erfüllt sind.

Was resultiert jedoch, wenn man auch über die betrachteten J Treatment-Stufen $X{=}x$, die jeweils als Referenz herangezogen werden, den theoretischen Mittelwert bildet? Dazu gehen wir im Weiteren davon aus, dass die unbedingten Treatment-Wahrscheinlichkeiten $P(X{=}x)$ für alle Treatment-Stufen bzw. für jede Klasse $X{=}x$ identisch sei, d. h. es gelte: $P(X{=}x) = J^{-1}$ für jeden Wert x von X. Der Erwartungswert $E(ACE_X)$ der durchschnittlichen kausalen Effekte ACE_x von x über alle Werte x von X ist dann:

$$
\begin{aligned}
E(ACE_X) &= \sum_{x=1}^{J} E(\delta_x) \cdot P(X{=}x) && (3.35)\\
&= J^{-1} \sum_{x=1}^{J} E(\delta_x) && P(X{=}x) = J^{-1}\\
&= J^{-1} \sum_{x=1}^{J} \Big[E(\tau_x) - J^{-1} \sum_{x'=1}^{J} E(\tau_{x'})\Big] && \text{[Gleichung 3.9]}\\
&= J^{-1} \sum_{x=1}^{J} E(\tau_x) - J^{-1} \sum_{x=1}^{J} \Big[J^{-1} \sum_{x'=1}^{J} E(\tau_{x'})\Big] && \text{[RR für Summen}^{33}\text{]}\\
&= J^{-1} \sum_{x=1}^{J} E(\tau_x) - J^{-1} \cdot J \cdot \Big[J^{-1} \sum_{x'=1}^{J} E(\tau_{x'})\Big] && \text{[RR für Summen}^{34}\text{]}\\
&= J^{-1} \sum_{x=1}^{J} E(\tau_x) - \Big[J^{-1} \sum_{x'=1}^{J} E(\tau_{x'})\Big]\\
&= J^{-1} \cdot \big[E(\tau_1) + \ldots + E(\tau_x) + \ldots + E(\tau_J)\big] -\\
&\quad J^{-1} \cdot \big[E(\tau_1) + \ldots + E(\tau_x) + \ldots + E(\tau_J)\big]\\
&= 0\,.
\end{aligned}
$$

Somit ist auch der theoretische Mittelwert der durchschnittlichen kausalen Effekte ACE_x von x über alle Werte x von X gleich null. Falls gilt $\forall x$ aus Ω_x: $P(X{=}x) = J^{-1}$ ist dies eine Eigenschaft der Effektparametrisierung, da jeder der auf diese Weise berechneten $(X{=}x)$-bedingten kausalen Effekte $CCE_{x;\,X=x}$ von x die Abweichung vom Mittelwert $J^{-1} \sum_{x'=1}^{J} E(\tau_{x'}\,|X{=}x)$ über alle Treatment-Stufen von X quantifiziert. Betrachtet man den Mittelwert der (gleichgewichteten) Abweichungen aller Werte von ihrem Mittelwert, so ist diese wiederum null.

[33] Rechenregel für Summen von Summen oder Differenzen: $\sum_i (a_i \pm b_i) = \sum_i a_i \pm \sum_i b_i$.
[34] Rechenregel für Summen gleicher Summanden: $\sum_i^n a = n \cdot a$.

3.5.4 Konsequenzen der Effektdefinition für die Interpretation

Die Definition der kausalen Effekte, auf deren Schätzung im Kontext von Vergleichsarbeiten gezielt wird, und die damit verbundenen Eigenschaften haben zwei zentrale Implikationen für deren inhaltliche Interpretation. Diese machen gleichfalls die Möglichkeiten und Grenzen derartiger Vergleiche deutlich. Diese Konsequenzen sind unabhängig von der Fairness-Problematik: Auch wenn wir in der Lage wären, kausale Effekte des Unterrichts auf die Testleistung der Schüler abzubilden, gelten die nachfolgenden Implikationen für die Effektschätzungen.

Eine erste Konsequenz: Wir betrachten niemals absolute Unterrichtseffekte.

Die Eigenschaft $E[\,E(\delta_{adj} \mid X)\,] = 0$ der adjustierten Effektfunktion, die auch für die entsprechende kausaltheoretische Größe zutrifft $[\,E(ACE_X) = 0\,]$, macht Folgendes deutlich: Wir betrachten nicht den absoluten Effekt eines Treatments im Vergleich zu *keinem* Treatment, sondern den Effekt eines Treatments im Vergleich zum Mittel über alle Treatments. Dieser Vergleich, den der *ACE on the treated* mit Effektparametrisierung beinhaltet, ist demnach nicht absolut, sondern normativ, wobei hier eine soziale Bezugsnorm verwendet wird. Der so definierte Effekt quantifiziert die jeweiligen Abweichungen – also das Mehr oder Weniger – im Vergleich zur mittleren Effektivität aller Treatments. „Schools or teachers are characterized as performing either above or below average compared with other units in the analysis [...]. In other words, estimates [...] have meaning only in comparison to average estimated effectiveness " (Braun, Chudowsky & Koenig, 2010, S. 24). Betrachten wir zwei Extremfälle, um die Konsequenzen einer derartigen Effektdefinition für die inhaltliche Interpretation zu verdeutlichen:

Beispiel 1: Nehmen wir zunächst an, der Unterricht in den betrachteten Klassen ist unwirksam, d. h. er hat keinerlei Effekte auf die Testleistung der Schüler: Unabhängig davon, ob ein Schüler zur Schule geht oder nicht, sei dessen Testleistung gleich. Dies gelte gleichermaßen für die Leistungsentwicklung des Schülers. Dann ist sowohl der durchschnittliche kausale Effekt des Unterrichts in einer Klassen x im Vergleich zu *keiner* Beschulung $ACE_{0x} = 0$, als auch der durchschnittliche kausale Effekt des Unterrichts in einer Klasse x im Vergleich zum Durchschnitt aller Klassen $ACE_x = 0$. Ebenso ist in diesem Fall der $(X{=}x)$-bedingte kausale Effekt des Unterrichts in einer

Klasse x (im Vergleich zum Durchschnitt aller Klassen) für die Schüler der jeweils betrachteten Klasse $CCE_{x;\,X=x} = 0$.

Beispiel 2: In einem zweiten Beispiel sei der Effekt des Unterrichts in einer Klasse im Vergleich zu keinem Unterricht jeweils positiv, jedoch sei der Unterricht aller Klassen in gleichem Ausmaß wirksam, d. h. $ACE_{0x} = ACE_{0x'} > 0$. Der positive Effekt des Unterrichts auf die Testleistung kann jedoch nicht detektiert werden, denn auch in diesem Fall ist sowohl der durchschnittliche kausale Effekt des Unterrichts in einer Klasse x im Vergleich zum Durchschnitt aller Klassen $ACE_x = 0$, als auch der $(X{=}x)$-bedingte kausale Effekt $CCE_{x;\,X=x} = 0$. Über die Effektivität des Unterrichts in diesem Beispiel kann lediglich gesagt werden, dass es zwischen den Klassen keine Unterschiede gibt; Aussagen über die *absolute* Effektivität des Unterrichts – und damit ohne soziale Bezugsnorm – sind hier jedoch nicht möglich.

Eine zweite Konsequenz: Rankings sind kontrafaktisch und somit obsolet.

Die auf die dargestellte Weise definierten klassenspezifischen kausalen Effekte, d. h. die $(X{=}x)$-bedingten kausalen Effekte $CCE_{x;\,X=x}$ des Unterrichts einer Klasse x, sind *zwischen* den Treatment-Bedingungen $X{=}x$ und $X{=}x'$ nicht ohne Weiteres miteinander vergleichbar. Ein Ranking dieser Effekte ist daher zunächst nicht bedeutsam. Das bereits vorgestellte Minimalbeispiel in Abbildung 3.3, bei dem wir lediglich von zwei Klassen bzw. zwei Treatment-Stufen $X = a$ und $X = b$ ausgehen, soll dies verdeutlichen. In diesem stark vereinfachten Szenario betrachten wir zwei Effekte, die in Abbildung 3.3 gelb markiert sind: (1) den $(X = a)$-bedingten kausalen Effekte $CCE_{a;\,X=a}$ des Unterrichts einer Klasse a und (2) den $(X = b)$-bedingten kausalen Effekte $CCE_{b;\,X=b}$ des Unterrichts einer Klasse b. Die $(X{=}x)$-bedingten kausalen Effekte $CCE_{x;\,X=x}$ sind jedoch nur *innerhalb* der $(X{=}x)$-Bedingung komparabel. Vergleichbar sind somit die Effekte $CCE_{a;\,X=a}$ und $CCE_{b;\,X=a}$. Alternativ kann man auch die Effekte $CCE_{a;\,X=b}$ und $CCE_{b;\,X=b}$ miteinander vergleichen. Ein Vergleich *zwischen* den Bedingungen $(X{=}x)$ – d. h. der Vergleich von $CCE_{a;\,X=a}$ mit $CCE_{b;\,X=b}$ – ist jedoch nicht ohne Weiteres bedeutsam, da es keine gemeinsame Referenz gibt. Diese sind nur dann vergleichbar, wenn sich die $(X{=}x)$-bedingten kausalen Effekte $CCE_{x;\,X=x}$ des Unterrichts einer Klasse x nicht vom durchschnittlichen kausalen Effekte ACE_x von x unterscheiden. Dies setzt voraus, dass es keine Interaktionen zwischen X und $\boldsymbol{Z}$ gibt, wovon im vorliegenden

Anwendungsfalls nicht ausgegangen werden kann.

3.6 Zusammenfassung

In diesem Kapitel habe ich die allgemeine stochastische Theorie kausaler Effekte nach Steyer et al. (2011) vorgestellt. Diese Theorie stellt eine Verallgemeinerung bisheriger kausaltheoretischer Ansätze in der Neyman-Rubin-Tradition (vgl. Neyman, 1923/1990; Rubin, 1974, 1977, 1978) dar, in der die stochastische Natur kausaler Zusammenhänge berücksichtigt wird. Die elementaren Bausteine zur Definition kausaler Effekte – die True-Outcome-Variablen τ_x – sind so definiert, dass sie von jeglichen Konfundierungen bereinigt sind. Zu diesem Zweck wird die umfassende Kovariate C_X betrachtet, die im gewissen Sinne *alle* potenziell konfundierenden Variablen umfasst.

Neben einer Gegenstandsbestimmung zwecks Abgrenzung des Kausalitätsbegriffes habe ich die mathematischen Grundkonzepte der Theorie vorgestellt, die den Kausalitätsraum definieren. Weiterhin wurden die Definitionen durchschnittlicher und bedingter kausaler Effekte dargestellt, die jeweils auf der True-Outcome-Variablen τ_x bzw. auf der Differenz $\delta_{xx'} \equiv \tau_x - \tau_{x'}$ der True-Outcome-Variablen basieren. Im Kontext von Schulleistungsuntersuchungen ist dabei insbesondere der *ACE on the treated* von Interesse. Anschließend wurde das Konzept der Unverfälschtheit eingeführt, welches das schwächste Kriterium zur Berechnung der theoretische Größen mittels empirisch schätzbarer Größen wie bspw. den *PFE* darstellt. Schließlich wurden vier hinreichende Bedingungen zur Identifikation kausaler Effekte – sog. *Kausalitätsbedingungen* – vorgestellt: die $\boldsymbol{Z}$-bedingte regressive Unabhängigkeit der True-Outcome-Variable τ_x von X ($\tau_x \vdash X \mid \boldsymbol{Z}, \forall x$), Rubins *Strong Ignorability* ($X \perp\!\!\!\perp \tau_x \mid \boldsymbol{Z}, \forall x$), die $\boldsymbol{Z}$-bedingte stochastische Unabhängigkeit der Treatment-Variable X und C_X ($X \perp\!\!\!\perp C_X \mid \boldsymbol{Z}$) sowie die $\boldsymbol{Z}$-bedingte regressive Unabhängigkeit der Outcome-Variable Y von C_X gegeben X ($Y \vdash C_X \mid X, \boldsymbol{Z}$). Die beiden Letzten sind dabei einer empirischen Prüfung – zumindest im Sinne der Falsifizierbarkeit – zugänglich und implizieren jeweils die bedingte regressive Unabhängigkeit $\tau_x \vdash X \mid \boldsymbol{Z}$. Diese wiederum ist die zentrale Bedingung im Rahmen der kausaltheoretischen Verortung statistischer Adjustierungsstrategien. Hier konnte gezeigt werden, dass die in Vergleichsarbeiten verwendeten Effektschätzungen nur unter sehr starken Annahmen den intendierten kausaltheoretischen Größen entsprechen, die in Anwendungen wenig plausibel sind. Im Kontext von Vergleichsarbeiten

kann somit bestenfalls von *faireren Vergleichen* ausgegangen werden. Zudem wurden anhand der Eigenschaften der auf diese Weise definierten Effekte die Möglichkeiten und Grenzen der inhaltlichen Interpretierbarkeit aufgezeigt.

You cannot put right by statistics what you have done wrong by design.

LIGHT & PILLEMER (1984)

4 Adjustierungsmodelle: Die Berechnung fairer(er) Vergleiche

Um die Fairness von unadjustierten und adjustierten Leistungsvergleichen im Kontext von Vergleichsarbeiten beurteilen zu können, bedarf es zunächst einer Systematisierung der verschiedenen Adjustierungsverfahren bzw. -modelle, die im Rahmen der Ergebnisrückmeldung von Vergleichsarbeiten Anwendung finden. Daher soll es in diesem Kapitel um folgende Fragen gehen: Welche Adjustierungsverfahren werden im Kontext von Vergleichsarbeiten angewendet, um potenziell faire Vergleiche anzustellen (Abschnitt 4.1.2)? Anhand welcher Kriterien lassen sich diese Adjustierungsverfahren charakterisieren (Abschnitt 4.1.3)? Wie unterscheiden sich die einzelnen Bundesländer in ihrem Vorgehen zur Berechnung fairer(er) Vergleiche und welche Gemeinsamkeiten gibt es (Abschnitt 4.1.4)? Und wie lassen sich diese Vorgehensweisen in den internationalen Kontext einordnen (Abschnitt 4.2)?

4.1 Systematisierung statistischer Adjustierungsverfahren im Kontext von Vergleichsarbeiten

Im Folgenden gebe ich einen Überblick über die gegenwärtige Praxis der einzelnen Bundesländer zur Berechnung und Rückmeldung von Referenzwerten vor dem Hintergrund fairer(er) Vergleiche. Dazu werden die derzeit verwendeten Adjustierungsverfahren beschrieben und in vier Kategorien eingeteilt, die anhand verschiedener Kriterien charakterisiert werden können. Anschließend werden die Vorgehensweisen der einzelnen Bundesländer bezüglich der Datenanalyse und Ergebnisrückmeldung diesen Kategorien statistischer Adjusterierungsverfahren zugeordnet. Die nachfolgende Dar-

stellung beruht im Wesentlichen auf Fiege (in Druck), Fiege et al. (2011) sowie Maaz, Trautwein und Dumont (2011).

4.1.1 Methodisches Vorgehen: Eine Quellenanalyse

Im Rahmen von Vergleichsarbeiten gibt es verschiedene statistische Adjustierungsstrategien, die die Problematik der Konfundierung, d. h. der Verzerrung der geschätzten Unterrichtseffekte durch Kovariaten, zu berücksichtigen suchen. Zumeist stehen in der entsprechenden Literatur jedoch die inhaltlichen Ergebnisse im Fokus der Betrachtung. Die konkrete Vorgehensweise zur Berechnung der Vergleichswerte und deren methodische Fundierung werden zumeist nur unzureichend oder überhaupt nicht dokumentiert. Aus diesem Grund wurde im Jahr 2009 eine umfassende Recherche durchgeführt, im Rahmen derer eine schriftliche und mündliche Befragung zur Praxis der Auswertung und Rückmeldung der einzelnen Landesinstitute und Ministerien stattfand. Auf Basis dieser Informationen wurde in Anlehnung an Nachtigall et al. (2008) und Fiege (2007) eine Systematik von Adjustierungsstrategien erstellt, der die verschiedenen Vergleichsarbeiten der einzelnen Bundesländer zugeordnet werden können (vgl. Fiege et al., 2011). Dabei unterschieden sich die statistischen Auswertungsstrategien im Rahmen von Vergleichsarbeiten nicht nur zwischen den einzelnen Bundesländern, sondern diese verändern sich auch im zeitlichen Verlauf. Im Jahr 2011 wurde auf Basis der Systematisierung von Fiege et al. (2011) eine erneute Recherche von Maaz et al. (2011) – ebenfalls mittels mündlicher und schriftlicher Befragung der Landesinstitute – durchgeführt. Die Zuordnung der einzelnen Bundesländer im Rahmen der vorliegenden Arbeit (vgl. Abschnitt 4.1.4) wurde basierend auf den Ergebnissen aus Maaz et al. (2011) sowie gegenwärtigen Entwicklungen aktualisiert und entspricht somit dem Stand des Schuljahres 2012/2013.

4.1.2 Kategorien von Adjustierungsstrategien

Das gemeinsame Merkmal der verschiedenen Adjustierungsstrategien ist, dass sich die Adjustierung jeweils auf die Berechnung des Vergleichs- bzw. Referenzwertes bezieht. So wird bspw. im Rahmen der klassenspezifischen Ergebnisrückmeldung des Projektes *Kompetenztest.de* der durchschnittliche Testwert einer Klasse (Mittelwerte der Testwerte der Schüler einer Klasse) einem adjustierten Vergleichswert – dem sog. *korrigierten*

Landesmittelwert – gegenübergestellt. Unabhängig von der Art der statistischen Adjustierung können dabei Lösungshäufigkeiten, Mittelwerte von Testleistungen, Kompetenzniveauverteilungen oder andere Verteilungskennwerte als Referenz dienen. Tabelle 4.1 stellt zusammenfassend die vier Kategorien von Referenzwerten dar, die sich im Rahmen der Adjustierung der Testergebnisse aus den bundeslandspezifischen Vergleichsarbeiten differenzieren lassen. Diese werden nachfolgend anhand prototypischer Beispiele näher erläutert.

Strategie I: Vergleich mit dem Landesmittelwert

Bei dieser ersten Strategie werden die entsprechenden Testwerte einer Klasse mit dem Landesmittelwert, d. h. dem arithmetischen Mittel der Testwerte aller Schüler der jeweiligen Klassenstufe innerhalb eines Bundeslandes, verglichen. Leistungsrelevante außerschulische Einflussgrößen des Lernens, auf die Schule und Lehrer keinen Einfluss haben, fließen hier nicht in die Berechnung des Referenzwertes ein. Eine statistische Adjustierung des Vergleichswertes findet demnach nicht statt. Das Ergebnis dieses Vorgehens sind unadjustierte und keine fairen Vergleiche – im Sinne der obigen Definition als kausale Unterrichtseffekte. Differenzen zum Landesmittelwert können lediglich im Rahmen einer sozialen Bezugsnorm interpretiert werden, ohne dabei belastbare Informationen über deren Ursachen zu liefern. Diese Auswertungs- und Rückmeldestrategie wird z. B. von Sachsen-Anhalt für die 3. Jahrgangsstufe (VERA 3) angewendet[1].

Strategie II: Vergleich mit einem subgruppenspezifischen Mittelwert

Die zweite Strategie im Rahmen der Datenanalyse der Testergebnisse aus Vergleichsarbeiten ist dadurch gekennzeichnet, dass eine *marginale* Adjustierung durchgeführt wird. Als Referenzwert wird der mittlere Testwert innerhalb einer bestimmten Subpopulation – also ein subgruppenspezifischer Mittelwert – berechnet und zum Vergleich herangezogen. So wird bspw. der durchschnittliche Testwert aus VERA 8 einer Gymnasialklasse in Brandenburg nicht dem Mittelwert aller Brandenburger Schüler, sondern dem durchschnittlichen Testwert aller Gymnasiasten der gleichen Klassenstufe gegenübergestellt. Neben der Schulform werden auch weitere Variablen im Rahmen der marginalen Adjustierung herangezogen. Auf diese Weise ist es möglich, den Auflösungs-

[1] Für weiterführende Informationen zu Vergleichsarbeiten in Sachsen-Anhalt siehe: `http://www.bildung-lsa.de`.

Tabelle 4.1: Kategorien von Adjustierungsstrategien in deutschen Vergleichsarbeiten

Strategie		Anmerkung	Beschreibung des Referenzwertes	Beispiel
I	Vergleich mit Landesmittelwert	Unadjustierte Vergleiche	Landesmittelwert	VERA 3 in Sachsen-Anhalt
II	Vergleich mit subgruppen-spezifischem Mittelwert	Marginale Adjustierung: Subklassifikation	Mittelwert innerhalb einer Subpopulation (bspw. Schulart, Geschlecht)	VERA 8 in Brandenburg
III	Vergleich mit ähnlichen (existierenden) Klassen	IIIa Standorttypen	Auswahl von Schulen des gleichen Standorttyps	Lernstand 8 in Nordrhein-Westfalen
		IIIb Belastungsindex	Auswahl von sechs Schulen mit ähnlichstem Belastungsindex	VERA 8 in Berlin
		IIIc Kontextgruppen	Auswahl von Schulen der gleichen Kontextgruppe	VERA 3 in Rheinland-Pfalz
IV	Vergleich mit Erwartungswert	IVa Outcome-Modellierung	Regressionsanalytisch vorhergesagter Wert unter Berücksichtigung verschiedener Kovariaten (Geschlecht, SES, Muttersprache etc.)	Lernstand 8 in Hessen
		IVb Outcome-Modellierung	+ fachspezifisches Vorwissen	Kompetenztest 8 in Thüringen

Anmerkungen. Kategorien in Anlehnung an Fiege et al. (2011).

grad bei der Ergebnisauswertung stetig weiter zu verfeinern. Ein Beispiel dafür sind die Ergebnisrückmeldungen zu VERA 8 in Brandenburg[2]: Hier werden die Klassen- und die Landesmittelwerte innerhalb einer Schulform separat für Mädchen zurückgemeldet, die anschließend miteinander verglichen werden können. Entsprechendes gilt für die jeweiligen Durchschnittswerte der Jungen.

Strategie III: Vergleich mit ähnlichen (existierenden) Klassen

Das zentrale Charakteristikum von Strategie III besteht darin, dass die Testwerte einer Klasse mit den Ergebnissen ähnlicher existierender Klassen verglichen werden. Die Ähnlichkeit von Klassen bezieht sich hier auf die soziale Zusammensetzung der Schülerschaft der betrachteten Klasse. Dies liegt darin begründet, dass der soziale Hintergrund der Schüler eine der wichtigsten Kovariaten im Schulleistungskontext darstellt (vgl. Baumert & Schümer, 2001). Im Rahmen dieser Strategie spricht man auch von *Kontextuierung* durch die Bildung konkreter Vergleichsgruppen (Wendt & Bos, 2011). Methodisch wird dies mittels der Konstruktion einer Kovariaten realisiert, welche die soziale Zusammensetzung bzw. die soziale Belastung einer Klasse quantifiziert (vgl. Bonsen et al., 2010). Zu diesem Zweck werden verschiedene Indikatoren des sozialen Hintergrundes der Schüler zu einem klassen- oder auch schulspezifischen Kennwert aggregiert. Zur Berechnung des Referenzwertes für eine spezifische Klasse werden dann ausschließlich die Testwerte aus Klassen mit der gleichen oder mit ähnlicher Ausprägung auf diesem Kennwert herangezogen. Fiege et al. (2011) differenzieren innerhalb dieser Strategie weiterhin drei Substrategien.

Strategie IIIa: Standorttypen. Bei der Ergebnisrückmeldung der Testergebnisse aus VERA 8 in Nordrhein-Westfalen[3] wird eine sog. *Standorttypisierung* bei der Erstellung fairer(er) Vergleiche verwendet. Diese Standorttypen charakterisieren einerseits regionale Merkmale des jeweiligen Schulstandortes und andererseits die soziodemographische Zusammensetzung der Schülerschaft innerhalb der verschiedenen Schulformen Hauptschule, Gesamtschule, Realschule und Gymnasium. Diese Zuweisung zu den verschiedenen Standorttypen erfolgte ursprünglich durch die Schulleitung der einzelnen

[2] Für weiterführende Informationen zu Vergleichsarbeiten in Brandenburg siehe: `http://www.isq-bb.de`.

[3] Für weiterführende Informationen zu Vergleichsarbeiten in Nordrhein-Westfalen siehe: `http://www.standardsicherung.nrw.de`.

Schulen selbst: Im Rahmen der Durchführung der jährlichen Lernstandserhebung wurden die Schulleiter der einzelnen Schulen aufgefordert, die Zuordnung ihrer Schule – und somit auch der getesteten Klassen – zu einem der Standorttypen vorzunehmen. Seit dem Schuljahr 2010/2011 wird diese jedoch vom Schulministerium basierend auf Daten der amtlichen Statistik vorgenommen. Dabei werden insgesamt fünf schulformübergreifende Standorttypen unterschieden. Bei der Zuordnung zu den Standorttypen werden neben dem Migrantenanteil der Schule gemäß der amtlichen Schulstatistik außerdem der Anteil der Empfänger von Arbeitslosengeld II unter 18 Jahren in der Umgebung der Schule berücksichtigt. Die Ergebnisrückmeldungen enthalten schließlich – neben den jeweiligen klassenspezifischen Ergebnissen – die Kompetenzniveauverteilungen aller Schüler aus Schulen des gleichen Standorttyps als Referenz.

Strategie IIIb: Belastungsindex. Auch in Hamburg wird bei der Ergebnisrückmeldung aus Vergleichsarbeiten, die seit dem Schuljahr 2012/2013 im Rahmen von KERMIT[4] erhoben werden, die Leistungsverteilung einer hinsichtlich der sozialen Zusammensetzung ähnlichen Schülerschaft als Referenz zurückgemeldet. Im Kontrast zu Strategie IIIa (Standorttypen) wird hier jedoch die jeweilige Referenzgruppe anhand eines metrischen Indikators für die soziale Belastung – dem sog. *Belastungsindex* oder *Sozialindex* einer Schule – bestimmt. Als klassen- und schulspezifische Referenz dienen „ . . . sechs Schulen der gleichen Schulform, deren Schülerschaft hinsichtlich der Bildungsnähe der Schülerfamilien ähnliche Voraussetzungen aufweist“ (Institut für Bildungsmonitoring und Qualitätsentwicklung, 2012, S. 6). Bei der Ergebnisrückmeldung wird dann zusätzlich zu den klassenspezifischen Ergebnissen die Leistungsverteilung dieser sechs Schulen berichtet. Die Konstruktion des Belastungsindex für die Schulen in Hamburg wird nicht in jedem Schuljahr erneut vorgenommen, sondern erfolgte im Rahmen der Hamburger Schulleistungsstudien KESS 4 und KESS 7 (Bos, Bonsen & Gröhlich, 2010; Bos & Pietsch, 2006). Anhand der dort erhobenen Fragebogendaten zum sozialen Hintergrund aus Schüler- und Elternbefragungen erfolgte eine empirische Analyse der sozialen Zusammensetzung der Schülerschaften an Hamburger Schulen in der Primar- und Sekundarstufe. Basierend auf diesen Daten wurden schließlich sowohl

[4] KERMIT (Kompetenzen ermitteln) ist ein System zur Erfassung der Kompetenzentwicklung der Schüler, das seit dem Schuljahr 2012/2013 in Hamburg etabliert wird. Dieses umfasst neben den Vergleichsarbeiten in Klassenstufe 3 und 8 zusätzlich die in Hamburg etablierten Lernausgangslagenerhebungen in den Klassenstufen 2, 5, 7 und 9. Für weiterführende Informationen zu Vergleichsarbeiten in Hamburg siehe: `https://www.lernstand.hamburg.de`.

die Hamburger Grundschulen als auch Schulen der Sekundarstufe I mittels eines Belastungsindex hinsichtlich der sozialen Zusammensetzung der Schülerschaft versehen (vgl. Bos, Gröhlich & Bonsen, 2010; Bos, Pietsch, Gröhlich & Janke, 2006).

Neben KERMIT in Hamburg lässt sich auch die Vorgehensweise in Berlin dieser Strategie IIIb zuordnen. Die Auswertung und Rückmeldung der Testergebnisse aus VERA 3 und VERA 8 in Berlin wird durch das *Institut für Schulqualität* (ISQ) durchgeführt[5]. Bis 2009 wurde hier in beiden Klassenstufen – ebenso wie in Brandenburg – Strategie II angewendet. Seit dem Schuljahr 2010/11 wird in Berlin hingegen die unterschiedliche soziale Zusammensetzung der Schülerschaft bei der Bestimmung einer Vergleichsgruppe für eine Schule berücksichtigt. Zu diesem Zweck wird auf zwei Merkmale der Schüler einer Klasse zurückgegriffen: (a) der Anteil der Schüler mit nichtdeutscher Herkunft und (b) der Anteil der Schüler, die von der Zuzahlung für Lernmittel befreit sind (Kuhl, Lenkeit, Pant & Wendt, 2011). Die Konstruktion der Referenzgruppe erfolgt schließlich in zwei Schritten: Zunächst werden alle Schulen entsprechend des Anteils der beiden oben genannten Merkmale in eine Rangreihe gebracht. In der Sekundarstufe I erfolgt dieser Schritt schulartspezifisch, d. h. ausschließlich Schulen der jeweils gleichen Schulart werden in eine Rangreihe gebracht. Zur Berechnung eines Referenzwertes für eine einzelne Schule werden anschließend die in dieser Rangreihe jeweils nächsten drei Schulen sowohl unterhalb als auch oberhalb der eigenen Position herangezogen. „Die Vergleichsgruppe [die zur Berechnung des Referenzwertes herangezogen wird] besteht damit insgesamt aus sechs Schulen mit einer sehr ähnlichen Zusammensetzung der Schülerschaft“ (Emmrich, Ernst, Harych & Wesselhöfft, 2012, S. 23). Im Unterschied zu der vergleichsweise komplexen und aufwendigen Vorgehensweise bei der Berechnung des Belastungsindex in Hamburg werden in Berlin der Anteil der Schüler mit nichtdeutscher Herkunft sowie der Anteil der Schüler mit Lehrmittelbefreiung als Indikatoren für den sozioökonomischen Status zugrunde gelegt.

Strategie IIIc: Kontextgruppen. Bei dieser Strategie handelt es sich um ein Analyseverfahren, das im Rahmen der Projektgruppe *Vergleichsarbeiten in der Grundschule* (Projektgruppe VERA) an der Universität Koblenz-Landau entwickelt wurde[6]. Dieses Verfahren wird in mehreren Bundesländern im Rahmen der Ergebnisrückmeldung bei VERA 3 angewendet, wobei zur Erstellung fairer(er) Vergleiche sog. *Kontextgruppen*

[5] Für weiterführende Informationen zu Vergleichsarbeiten in Berlin siehe: `http://www.isq-bb.de`.
[6] Für weiterführende Informationen zur Projektgruppe VERA siehe: `http://www.projekt-vera.de`.

gebildet werden (Isaac & Hosenfeld, 2008). Methodisch handelt es sich hier um ein zweistufiges Verfahren: In einem ersten Schritt erfolgt die Berechnung von Kontextwerten anhand einer bundeslandspezifischen repräsentativen Zufallsstichprobe, der sog. *Zentralstichprobe*. Die Kontextwerte werden im Rahmen eines regressionsanalytischen Mehrebenen-Ansatzes (Hierarchisch Lineare Modellierung, HLM) ermittelt. Zu diesem Zweck wird der Gesamtleistungswert – d. h. der Mittelwert der Schülerleistungen in den beiden getesteten Fächern Deutsch und Mathematik – durch verschiedene Prädiktoren vorhergesagt. Die dazu verwendeten leistungsrelevanten Kovariaten sind Prädiktoren auf Schüler- und auf Klassenebene (z. B. das Geschlecht, der klassenspezifische Anteil von Kindern aus Familien mit Arbeitslosigkeit oder der klassenspezifische Anteil von Jungen). Ziel dieser Analyse ist es, möglichst viel Varianz aufzuklären. Nur solche Prädiktoren, die sich als statistisch signifikant erweisen, fließen in die weiteren Berechnungen ein. Die für die einzelnen Klassen der Zentralstichprobe aufgrund dieser Prädiktoren vorhergesagten Leistungswerte sind die Kontextwerte. Diese Kontextwerte sind die unter Berücksichtigung leistungsrelevanter außerschulischer Einflussgrößen zu erwartenden Leistungswerte der Schülerschaft, wobei hier u. a. auch die soziale Zusammensetzung einer Klasse berücksichtigt wird. Anschließend wird die Verteilung der berechneten Kontextwerte in drei Gruppen eingeteilt: Kontextgruppe I und III enthält jeweils 25% aller Klassen mit den niedrigsten bzw. höchsten Kontextwerten. Kontextgruppe II schließt hingegen die mittleren 50% der Klassen ein. Im zweiten Analyseschritt erfolgt die Zuweisung aller Klassen eines Bundeslandes zu den drei Kontextgruppen. Die Zuordnung basiert einerseits auf freiwilligen Lehrerangaben zur Einschätzung des sozialen Hintergrundes der eigenen Klasse und andererseits den Schülerstammdaten. Die klassenspezifischen Ergebnisrückmeldung enthalten schließlich die Kompetenzniveauverteilung aller Klassen aus der eigenen Kontextgruppe als Referenz.

Strategie IV: Vergleich mit einem Erwartungswert

Im Rahmen von Strategie IV werden Vergleiche nicht bezüglich der Testleistung tatsächlich existierender Klassen vorgenommen, sondern bezüglich eines regressionsanalytisch vorhergesagten Wertes bzw. eines klassenspezifischen Erwartungswertes. Dieser klassenspezifische Erwartungswert stellt den für eine Klasse mit ähnlichen Kontextbedingungen hinsichtlich relevanter Schülermerkmale zu erwartenden durchschnittlichen Testwert dar. Innerhalb dieser Adjustierungsstrategie können wiederum zwei Vorge-

hensweisen differenziert werden. Diese unterscheiden sich darin, ob zusätzlich zu den bereits erwähnten Kovariaten gleichfalls das fachspezifische Vorwissen der Schüler berücksichtigt wird.

Strategie IVa. Dieser Strategie lässt sich das vom Projekt *Kompetenztest.de* an der Friedrich-Schiller-Universität Jena entwickelte Adjustierungsverfahren (Nachtigall & Kröhne, 2006; Nachtigall et al., 2008) zuordnen[7]. Dieses wird u. a. im Rahmen der Thüringer Kompetenztests in Klassenstufe 3 oder auch bei den Hessischen Lernstandserhebungen in Klassenstufe 8 angewendet. Dabei wird zunächst für jeden Schüler einer konkreten Klasse ein Erwartungswert – d. h. ein vorhergesagter Mittelwert – bestimmt. Diese schülerspezifischen Erwartungswerte werden über die Zellenmittelwerte einer multifaktoriellen Varianzanalyse (ANOVA-Zellenmittelwertemodell) geschätzt (vgl. Nachtigall et al., 2008). Die abhängige Variable ist die Testwertvariable, deren Werte die Kompetenztestergebnisse der einzelnen Schüler sind. Die verschiedenen Faktoren sind die Kovariaten wie bspw. die Schulart oder das Geschlecht. Der Erwartungswert eines Schülers ist demnach der Mittelwert der Testleistungen aller Schüler mit der gleichen Kovariatenkonstellation, d. h. den gleichen Ausprägungen auf den berücksichtigten Kovariaten. Der adjustierte Referenzwert einer Klasse, der sog. *korrigierte Landesmittelwert*, ist schließlich der Mittelwert der geschätzten schülerspezifischen Erwartungswerte der jeweils betrachteten Klasse.

Strategie IVb. Eine Besonderheit der Vorgehensweise im Projekt *Kompetenztest.de* besteht im Hinblick auf die Adjustierung bei den Thüringer Kompetenztestergebnissen aus Klassenstufe 8. Hier wird zusätzlich zu den bereits erwähnten Kovariaten auch das fachspezifische Vorwissen, d. h. das Testergebnis aus dem Kompetenztest der jeweils früheren Klassenstufe berücksichtigt: Neben den Schülerstammdaten (z. B. das Geschlecht der Schüler) und dem sozioökonomischen Hintergrund wird zusätzlich der Vortestwert der Schüler aus den Kompetenztests der früheren Klassenstufe in die Analyse einbezogen. Für die Schüler aus Klassenstufe 8 werden somit deren Testergebnisse aus Klassenstufe 6 verwendet. Dieser Vortestwert ist ein Indikator für das fachspezifische Vorwissen eines Schülers, das wiederum eine weitere zentrale Determinante der Schülerleistungen darstellt (vgl. Hedges & Hedberg, 2007; Nachtigall et al., 2008;

[7] Für weiterführende Informationen zum Projekt *Kompetenztest.de* siehe: `http://www.kompetenztest.de`.

Renkl, 1996; Schrader & Helmke, 2008). Das Vorwissen einer Person umfasst ihre Kenntnisse (deklaratives Wissen) und Fertigkeiten (prozedurales Wissen) in einem bestimmten Gegenstandsbereich (Domäne) bzw. Schulfach (vgl. Renkl, 1996). Die Verwendung des fachspezifischen Vorwissens der Schüler bei der Datenanalyse setzt voraus, dass längsschnittliche Daten zur Verfügung stehen.

Gemeinsames Merkmal dieser beiden Strategien – d. h. sowohl von Strategie IVa als auch von Strategie IVb – ist die Berechnung eines Erwartungswertes, der den zu erwartenden durchschnittlichen Testwert einer Klasse mit vergleichbarer Zusammensetzung hinsichtlich diverser Kovariaten auf Schüler- oder auch Klassenebene repräsentiert. Dabei sei an dieser Stelle nochmals betont, dass die zusätzliche Verwendung des fachspezifischen Vorwissens kein genuines Merkmal der Strategie IV ist, sondern ein Spezifikum des Vorgehens im Projekt *Kompetenztest.de* im Rahmen der Thüringer Kompetenztests der Klassenstufe 8 (Strategie IVb).

4.1.3 Kriterien zur Bewertung der Adjustierungsstrategien

Nach Fiege et al. (2011) können die verschiedenen Adjustierungsstrategien im Hinblick auf mindestens drei Kriterien charakterisiert werden. Diese stehen in engem Zusammenhang und bilden so eine Triade zur Beurteilung von Adjustierungsverfahren (vgl. auch Fiege, in Druck). Die drei Kriterien sind (1) die Fairness, (2) die Testökonomie sowie (3) die Komplexität des im Rahmen der Datenanalyse verwendeten statistischen Modells. Die beiden zuletzt genannten – Testökonomie und Modellkomplexität – umfassen zwei Kriterien, die insbesondere die Praktikabilität des eingesetzten statistischen Verfahrens charakterisieren.

(1) *Fairness:*
In Kapitel 3 wurde der Fairness-Begriff bereits ausführlich vor dem Hintergrund einer allgemeinen stochastischen Theorie kausaler Effekte betrachtet und analysiert. Die Fairness einer Adjustierungsstrategie bezieht sich somit auf die Definition fairer, d. h. kausal interpretierbarer Vergleiche. Eine wesentliche Folgerung aus dieser Definition ist, dass erst durch die Berücksichtigung von Kovariaten – also außerschulischer Einflussgrößen des Lernens, auf die der Lehrer einer Klasse keinen Einfluss hat – potenziell faire Vergleiche möglich sind. Eine notwendige wenngleich auch nicht hinreichende Bedingung fairer Vergleiche im Sinne der

kausalen Definition ist, dass *alle relevanten* Kovariaten in die Analyse einbezogen werden müssen. Dies ist im schulischen Kontext in der Regel nicht realisierbar. So können allein aus testökonomischen Gründen nicht sämtliche relevanten außerschulischen Einflussfaktoren des schulischen Lernens erhoben werden. Ein realistisches Ziel hingegen ist die Identifikation und Berücksichtigung der für schulische Leistungen zentralen Kovariaten wie bspw. des sozioökonomischen Status, des Geschlechts und des fachspezifischen Vorwissens. Die Auswahl der Kovariaten muss dabei stets auch von der Verteilung der relevanten Kovariaten in der jeweils betrachteten Population abhängen. „In diesem Sinne stellen Adjustierungsverfahren in der Praxis empirischer Bildungsforschung in der Regel eine Annäherung an faire Vergleiche dar“ (Fiege et al., 2011, S. 138), so dass wir maximal von faireren Vergleichen sprechen können. Hier unterscheiden sich die Adjustierungsstrategien hinsichtlich der Art (Welche Kovariaten werden berücksichtigt?) und Anzahl (Wie viele Kovariaten fließen in die Analyse ein?) der berücksichtigen Kovariaten. Werden keine Kovariaten berücksichtigt, so ist der Vergleich in diesem Sinne als *unfair* zu erachten.

(2) *Testökonomie:*
Ein zweites Kriterium ist die Testökonomie (vgl. Moosbrugger & Kelava, 2007). Diese bezieht sich auf die Erfassung der in der Analyse berücksichtigten Kovariaten. So werden in einigen Bundesländern Daten der amtlichen Statistik als Informationsquelle genutzt. Hingegen werden in anderen Ländern zusätzliche Fragebögen für Schüler und Lehrer eingesetzt, um Informationen hinsichtlich relevanter Kovariaten zu erheben. Letztere Vorgehensweise setzt nicht nur die Motivation, sondern auch zeitliche Ressourcen der Schüler und Lehrer voraus – zusätzlich zu dem ohnehin schon recht aufwändigen Testverfahren. Die Testökonomie adressiert somit die Frage, ob die relevanten außerschulischen Einflussfaktoren des schulischen Lernens sparsam und ohne nennenswerte zusätzliche Kosten wie Zeit, Geld oder andere Ressourcen erfasst werden können.

(3) *Komplexität des statistischen Modells:*
Auch hinsichtlich des methodischen Vorgehens bei der statistischen Datenanalyse lässt sich ein Sparsamkeitskriterium differenzieren, das dem sog. *Parsimonitätsprinzip* entspricht. Hierbei geht es um die Komplexität des Modells, also die Frage, ob zur Berechnung des adjustierten Vergleichswertes komplexe statis-

tische Modelle angewendet werden. Die Modelle sollten dabei so komplex wie nötig sein, d. h., sie sollen die tatsächlich bestehenden Zusammenhänge der Variablen adäquat abbilden können. Dies ist eine zweite notwendige Bedingung für faire Vergleiche (vgl. Kapitel 3). Die Modelle sollten aber auch so einfach wie möglich sein, da in komplexen statistischen Modellen mehr Parameter geschätzt werden müssen. Dadurch steigen die Anforderungen an die zugrundeliegenden Daten, denn um diese Parameter schätzen zu können, sind z. T. zusätzliche Annahmen bzw. größere Stichprobenumfänge nötig. Die Beobachtungsanzahl ist im schulischen Kontext jedoch natürlich begrenzt durch die Anzahl der Schüler einer Klasse, einer Schule bzw. eines Bundeslandes. Zudem sind einfachere, weniger komplexe statistische Modelle leichter zu verstehen – ein nicht zu vernachlässigender Aspekt der Ergebnisdissemination: Für die Rezeption und Verwertung der Ergebnisse im Rahmen von Unterrichtsentwicklungsmaßnahmen kann dies von Vorteil sein, da Transparenz bezüglich der Datenanalysen potenziell zu mehr Verständnis und erhöhter Akzeptanz führt. Demgemäß spricht Briggs (2008) die Empfehlung aus: „When stakes are low, it may be more valueable to use less data with a simpler model such that the process becomes more transparent to stakeholders“ (S. 11).

Auch Kuhl et al. (2011) verwenden ähnliche Bewertungskriterien. Die Autoren charakterisieren unterschiedliche Adjustierungsmodelle hinsichtlich einer Kosten- sowie einer Nutzendimension. Die *Kosten* werden über den Aufwand der Datenbeschaffung operationalisiert. Dies entspricht der Testökonomie bei Fiege et al. (2011): Je ökonomischer die Datenerhebung, d. h. je praktikabler das Testverfahren, desto geringer sind die Kosten einer konkreten Adjustierungsstrategie. Der *Nutzen* eines Adjustierungsmodells hingegen wird bei Kuhl et al. (2011) über den Anteil erklärter Varianz quantifiziert. Implizit werden hier Nutzen und Fairness gleichgesetzt. Dabei wird – ebenso implizit und nicht explizit – die Annahme gemacht, dass durch die Aufnahme weiterer Kovariaten in das Adjustierungsmodell, die zusätzlich Varianz in den Leistungsdaten aufklären, die adjustierten Effektschätzungen verbessert werden und man somit zu fairen oder zumindest zu faireren Vergleichen kommt. „Optimal im Sinne der Kosten-Nutzen-Abwägung wäre demnach ein Modell, bei dem der Aufwand der Datenerhebung möglichst gering und zugleich der Anteil aufgeklärter Leistungsvarianz möglichst groß ist“ (Kuhl et al., 2011, S. 239). Bei der (Weiter-)Entwicklung von Adjustierungsstrategien sollte das Ziel

somit in der Maximierung der Fairness bei gleichzeitiger Maximierung der Praktikabilität liegen.

Mit Blick auf die Kategorisierung der Adjustierungsstrategien (vgl. Abschnitt 4.1.2 und Tabelle 4.1) lässt sich schließlich feststellen, dass die Fairness mit steigender Ordnungszahl zunimmt. Gleichzeitig nimmt die Praktikabilität (Testökonomie und Modellkomplexität) tendenziell ab. Ein Beispiel soll dies verdeutlichen: Im Gegensatz zu Strategie I, die mit unadjustierten Vergleichen weder faire noch fairere Vergleiche liefert, ist Strategie II als fairer zu beurteilen, da relevante Kovariaten (bspw. Geschlecht, Schulart) berücksichtigt werden. Ein weiterer Zugewinn hinsichtlich der Fairness kann durch die Strategien III bzw. IV erzielt werden, die jedoch beide weniger praktikabel sind: Diese Strategien kommen nicht mehr mit Schülerstammdaten aus, sondern erfordern die Zusatzerhebung weiterer Kovariaten. Im strengen Sinne, d. h. mit dem Ziel kausal interpretierbarer Vergleiche, liefert jedoch keine der Adjustierungsstrategien den *fairen* Vergleich (vgl. Kapitel 3).

4.1.4 Adjustierungsstrategien in den Bundesländern

Nachdem bisher die verschiedenen Adjustierungsstrategien vorgestellt, kategorisiert sowie hinsichtlich der Kriterien Fairness und Praktikabilität beurteilt wurden, werde ich nachfolgend die Frage beantworten, in welchem Bundesland welche Adjustierungsstrategie angewendet wird. Tabelle 4.2 und Tabelle 4.3 zeigt die Zuordnung der Bundesländer zu den einzelnen Kategorien von Adjustierungsstrategien aus Abschnitt 4.1.2, die im Kontext der Ergebnisauswertung und -rückmeldung der Testergebnisse aus Vergleichsarbeiten der Klassenstufen 3 und 8 angewendet werden (vgl. Fiege, in Druck; Fiege et al., 2011; Maaz et al., 2011).

Neben der Zuordnung der Bundesländer wird zusätzlich in beiden Tabellen markiert, falls im Jahr 2009 ein anderes Verfahren verwendet wurde als im Jahr 2011 oder später. Dadurch wird gekennzeichnet, falls sich innerhalb von zwei Jahren Veränderungen hinsichtlich des Vorgehens bei der Datenanalyse und Rückmeldung der Testergebnisse vollzogen haben. Damit wird noch einmal deutlich, dass weder die Kategorisierung noch die Zuordnung der Bundesländer feststehende Eigenschaften von Vergleichsarbeiten sind, sondern insbesondere die Datenanalyse und Ergebnisrückmeldungen einem steten Entwicklungsprozess unterliegen.

Tabelle 4.2: Adjustierungsstrategien im Kontext von Vergleichsarbeiten der 3. Jahrgangsstufe (VERA 3) in den einzelnen Bundesländern

Bundesland	Adjustierungsstrategie						
	I	II	III			IV	
			IIIa	IIIb	IIIc	IVa	IVb
Baden-Württemberg	×				★		
Bayern	×						
Berlin		★		×			
Brandenburg		×					
Bremen					×		
Hamburg				×			
Hessen						×	
Mecklenburg-Vorpommern					×		
Niedersachsen					×		
Nordrhein-Westfalen					×		
Rheinland-Pfalz					×		
Saarland					×		
Sachsen						×	
Sachsen-Anhalt	×						
Schleswig-Holstein					×		
Thüringen						×	

Anmerkungen. Die Adjustierungsstrategien I bis IV basieren auf der Systematik nach Fiege et al. (2011). Die Zuordnungen der Bundesländer sind mit einem × gekennzeichnet und basieren auf einer schriftlichen Befragung der Landesinstitute im Jahr 2011 (vgl. Maaz et al., 2011) sowie einer aktuellen Recherche im Schuljahr 2012/2013. Falls im Jahr 2009 eine davon abweichende Strategie verwendet wurde, ist dies mit einem ★ gekennzeichnet.

Tabelle 4.3: Adjustierungsstrategien im Kontext von Vergleichsarbeiten der 8. Jahrgangsstufe (VERA 8) in den einzelnen Bundesländern

Bundesland	Adjustierungsstrategie						
	I	II	III			IV	
			IIIa	IIIb	IIIc	IVa	IVb
Baden-Württemberg [a]		×					
Bayern		×					
Berlin		★		×			
Brandenburg		×					
Bremen		★			×		
Hamburg				×			
Hessen						×	
Mecklenburg-Vorpommern						×	
Niedersachsen					×	★	
Nordrhein-Westfalen			×				
Rheinland-Pfalz		×					
Saarland		×					
Sachsen						×	
Sachsen-Anhalt		×					
Schleswig-Holstein		×					
Thüringen							×

Anmerkungen. Die Adjustierungsstrategien I bis IV basieren auf der Systematik nach Fiege et al. (2011). Die Zuordnungen der Bundesländer sind mit einem × gekennzeichnet und basieren auf einer schriftlichen Befragung der Landesinstitute im Jahr 2011 (vgl. Maaz et al., 2011) sowie einer aktuellen Recherche im Schuljahr 2012/2013. Falls im Jahr 2009 eine davon abweichende Strategie verwendet wurde, ist dies mit einem ★ gekennzeichnet.

[a] Die Entwicklung der Items und die Zusammenstellung der Tests (VERA 3 und VERA 8) erfolgen am IQB unter Mitwirkung aller Bundesländer. Baden-Württemberg ist das einzige Bundesland, das nicht an VERA 8 im Bundesverband teilnimmt, sondern eigene Vergleichsarbeiten für die Sekundarstufe I auf Basis der baden-württembergischen Bildungsstandards entwickelt. Die Tests beziehen sich auf zweijährige Bildungsabschnitte und werden zu Beginn der 9. Jahrgangstufe durchgeführt (vgl. Wacker & Kramer, 2012).

Veränderungen der Adjustierungsstrategie über die Zeit

Summa summarum zeigt sich ein weitestgehend stabiles Bild über die Zeit: Zwischen 2009 und 2013 gibt es nur wenige Veränderungen der Adjustierungsstrategien innerhalb der einzelnen Bundesländer.

In Baden-Württemberg wurden die Ergebnisse aus VERA 3 zwischen 2009 und 2011 von der Projektgruppe VERA an der Universität Koblenz-Landau ausgewertet (vgl. Tabelle 4.2). Die Ergebnisrückmeldungen enthielten auf freiwilliger Basis faire(re) Vergleiche (Strategie IIIc). Seit dem Schuljahr 2011/2012 wird die Datenanalyse und Erstellung der Ergebnisrückmeldungen hingegen über ein landeseigenes VERA-Online-Portal realisiert, wobei hier lediglich ein landesweiter Vergleichswert (Landesmittelwert) zur Verfügung gestellt wird (Strategie I). Ein weiterer Wechsel der Vorgehensweise hinsichtlich von Vergleichsarbeiten in der 3. Jahrgangsstufe wurde in Berlin vorgenommen: Während im Jahr 2009 noch Strategie II angewendet wurde, wird seit dem Schuljahr 2010/2011 die unterschiedliche Zusammensetzung der Schülerschaft bei der Berechnung einer Vergleichsgruppe für eine Schule berücksichtigt (Strategie IIIb; vgl. Abschnitt 4.1.2).

Im Rahmen von Vergleichsarbeiten der Klassenstufe 8 (VERA 8) gab es lediglich in Berlin, Bremen und Niedersachsen Veränderungen bezüglich der Adjustierungsstrategie (vgl. Tabelle 4.3): In Berlin wird nun – wie auch in Klassenstufe 3 – anstelle von Strategie II Strategie IIIb angewendet (vgl. Abschnitt 4.1.2). Niedersachsen ließ die Daten aus den Vergleichsarbeiten in Klassenstufe 8 im Jahr 2009 durch das Projekt *Kompetenztest.de* an der Friedrich-Schiller-Universität Jena auswerten, bei dem Strategie IVa angewendet wird. Ab dem Jahr 2011 hingegen wurde die Auswertung durch die Projektgruppe VERA an der Universität Koblenz-Landau durchgeführt. Hier wird Strategie IIIc angewendet, sofern im Rahmen der jährlichen Erhebung der Vergleichsarbeiten die entsprechenden Informationen (d. h. Kovariaten) erhoben werden. Auch Bremen ließ im Jahr 2011 die Auswertung der Testergebnisse aus den Vergleichsarbeiten der 8. Jahrgangsstufe durch die Projektgruppe VERA an der Universität Koblenz-Landau durchführen, wohingegen im Jahr 2009 die Auswertung noch auf Strategie II beruhte.

Ein Vergleich zwischen den Bundesländern

Vergleicht man die Adjustierungsstrategien der einzelnen Bundesländer in Klassenstufe 3 (VERA 3; vgl. Tabelle 4.2), so fällt auf, dass sieben der 16 Bundesländer (Bremen, Mecklenburg-Vorpommern, Niedersachsen, Nordrhein-Westfalen, Rheinland-Pfalz, Saarland und Schleswig-Holstein) Strategie IIIc anwenden. Dies ist darauf zurückzuführen, dass diese sieben Bundesländer die Auswertung durch die Projektgruppe VERA an der Universität Koblenz-Landau durchführen lassen. Weiterhin wird Strategie IVa von den drei Bundesländern Hessen, Sachsen und Thüringen verwendet. Auch diese Länder befinden sich in einem Auswertungsverbund: Die Auswertung der Testergebnisse und Erstellung der Rückmeldungen erfolgt durch das Projekt *Kompetenztest.de* an der Friedrich-Schiller-Universität Jena. Die Auswertung der Testergebnisse aus Berlin und Brandenburg erfolgt ebenfalls durch dieselbe Institution – das ISQ, das für die Daten aus Brandenburg Strategie II anwendet. Für die Daten aus Berliner Schulen wird jedoch – wie in Abschnitt 4.1.2 ausführlich dargestellt – Strategie IIIb verwendet. Auf diese Weise wird versucht, der im Vergleich zu Brandenburg heterogeneren soziodemographischen Zusammensetzung der Schülerschaft bspw. hinsichtlich des Anteils von Schülern mit Migrationshintergrund Rechnung zu tragen. Dies macht deutlich, dass die Wahl eines Adjustierungsverfahrens auch von der Verteilung der relevanten Kovariaten in der jeweils betrachteten Population abhängen muss. So müssen also stets auch die spezifischen Gegebenheiten – bezogen auf die Zusammensetzung der Schülerschaft hinsichtlich relevanter Kovariaten – eines Bundeslandes bei der Wahl der Adjustierungsstrategie besondere Berücksichtigung finden. Dies stellt gleichsam eine zusätzliche Herausforderung im Rahmen der Formulierung allgemein verbindlicher Standards bzw. Richtlinien bei der Erstellung fairer(er) Vergleiche dar.

Für die Adjustierungsstrategien in Klassenstufe 8 (VERA 8; vgl. Tabelle 4.3) zeigt sich ein heterogeneres Bild beim Vergleich der Bundesländer. Zwar erfolgt die Auswertung der Bundesländer gemäß dem Vorgehen aus Kategorie IVa (Hessen, Mecklenburg-Vorpommern, Sachsen) sowie IVb (Thüringen) gleichfalls durch das Projekt *Kompetenztest.de* an der Friedrich-Schiller-Universität Jena. Der Auswertungsverbund der Länder, die die Auswertung durch die Projektgruppe VERA an der Universität Koblenz-Landau durchführen lassen, besteht hier jedoch nur aus vier Bundesländern (Bremen, Niedersachsen, Rheinland-Pfalz, Saarland). Zusätzlich gibt es auch innerhalb dieser vier Bundesländer Unterschiede im Datenanalyse- und Rückmeldeformat in Abhängig-

keit davon, welche Kovariaten für die Datenanalyse zur Verfügung stehen. Der Grund dafür ist, dass die Anwendung von Strategie IIIc nur möglich ist, wenn zusätzliche Informationen zu außerschulischen Einflussgrößen des Lernens (z. B. der soziale Hintergrund der Familie) in der jährlichen Erhebung der Vergleichsarbeiten erfasst werden.

4.2 Einordnung in den internationalen Kontext: Ein Vergleich mit den USA und England

Im vorangegangenen Abschnitt wurden die in den Bundesländern verwendeten Vorgehensweisen zur Berücksichtigung von außerschulischen Einflussgrößen des Lernens beim Vergleich von Testergebnissen aus deutschen Vergleichsarbeiten vorgestellt. Dabei lassen sich vier verschiedene Adjustierungsstrategien differenzieren, die sich hinsichtlich der Testökonomie, Modellkomplexität und der Fairness der resultierenden Vergleiche unterscheiden. Doch wie lassen sich diese Vorgehensweisen im internationalen Kontext verorten? Gibt es in anderen Ländern ähnliche Vorgehensweisen bzw. Adjustierungsmodelle, die im Rahmen der Qualitätsentwicklung und -sicherung der jeweiligen Bildungssysteme Anwendung finden? Inwiefern gibt es Gemeinsamkeiten und wo liegen Unterschiede? Gibt es in anderen Ländern erweiterte Modelle, die geeigneter sind zur Berechnung fairer(er) Vergleiche?

Exemplarisch werden nachfolgend zwei Länder betrachtet, die jeweils eine ausgeprägte Testkultur im schulischen Bereich (Oelkers & Reusser, 2008) entwickelt haben: die USA und England. Der Paradigmenwechsel von der Input- und Prozessorientierung hin zur Evaluation des Outputs von Schule vollzog sich in anderen Ländern früher als in Deutschland (vgl. Abschnitt 2.2). Testbasierte Evaluationssysteme zur Schul- und Unterrichtsentwicklung sowie zur öffentlichen Rechenschaftslegung – sog. *Educational-Accountability-Systeme* (vgl. z. B. Ryan & Shepard, 2008) – sind in einigen Ländern seit vielen Jahren Grundlage bildungspolitischer Reformen, schulischer Qualitätssicherung oder elterlicher Schulwahl. Jedoch soll an dieser Stelle kein detaillierter historischer Abriss zur Entwicklung der Schulsysteme dieser Länder oder deren Qualitätssicherungssysteme erfolgen. Hierzu verweise ich den interessierten Leser auf die entsprechende Literatur (z. B. Braun et al., 2010; Klieme, Döbert et al., 2007; Lind, 2009; Oelkers & Reusser, 2008; van Ackeren, 2002, 2003b, 2003a sowie die dort zitierte Literatur). Im Fokus des Interesses stehen nachfolgend die im Rahmen der nationalen

Leistungstests dieser Ländern verwendeten Adjustierungsverfahren und insbesondere deren Parallelen zu den in Deutschland verwendeten Verfahren.

4.2.1 State Achievement Tests in den USA

Bereits seit den 1950er Jahren ist die testbasierte Evaluation des Outputs von Schule in den USA ein verbreitetes und anerkanntes Mittel für Selektionsentscheidungen sowie zur Rechenschaftslegung und Verbesserung des Bildungssystems (Linn, 2000). Diese Entwicklung erreichte ihren Höhepunkt unter George W. Bush mit dem *No Cild Left Behind Act* (NCLB) aus dem Jahr 2001 – einem Bildungsgesetz zur Qualitätsverbesserung öffentlicher Schulen (NCLB, 2002). Im NCLB-Gesetz wird die Zuweisung staatlicher Fördermittel zu den einzelnen Schulen geregelt, indem die Finanzierung der Schulen von der Evaluation schulischer Leistung abhängig gemacht wird. Das bedeutet, dass alle öffentlichen, mit Bundesmitteln geförderten Schulen jährlich standardisierte Tests, sog. *state achievement tests*, in den Jahrgangstufen 3 bis 8 sowie einmal in der High School in den Fächern Mathematik, Englisch und Naturwissenschaften durchführen müssen. Anhand ihrer Testwerte werden die Schüler hinsichtlich des Erreichens verschiedener Kompetenzstufen klassifiziert, z. B. in „unsatisfactory“, „proficient“ und „advanced“. Die Entwicklung von Leistungsstandards (*performance standards*) sowie die Definition der einzelnen Kompetenzstufen liegt jedoch jeweils in der Verantwortung der einzelnen Bundesstaaten. Das NCLB-Gesetz legt das gemeinsame Ziel fest, dass bis zum Jahr 2014 alle Schüler eines Staates mindestens das Kompetenzniveau „proficient“ erreichen müssen. Zusätzlich muss jedoch in jedem Jahr bis 2014 ein bestimmter, jährlich steigender Prozentsatz der Schüler als „proficient“ klassifiziert werden. Dieses Ziel, der sog. *adequate yearly progress* (AYP), wird für die gesamte Schülerpopulation eines Bundesstaates als auch innerhalb bestimmter Subgruppen (bspw. nach Geschlecht, SES, Ethnizität etc.) wiederum durch die einzelnen Bundesstaaten selbst festgelegt. Wird der AYP innerhalb einer Schule nicht erreicht, ist diese „in need of improvement“, wobei dies in der öffentlichen Diskussion in den Medien häufig mit dem Label „failed (gescheitert)“ gekennzeichnet wird. Als Konsequenz drohen dieser Schule verschiedene Sanktionen: Nach zwei Jahren des Verfehlens vom AYP muss diese ihren Schülern ermöglichen, in eine andere Schule zu wechseln und die dabei anfallenden Transportkosten übernehmen. Wird der AYP von einer Schule in drei aufeinanderfolgenden Jahren nicht erreicht, muss diese u. a. zusätzliche Fördermaßnahmen (*supplemental educatio-*

nal services) anbieten. Nach fünf Jahren droht der Schule sogar die Auflösung (vgl. z. B. Bracey, 2005; Briggs, 2008; Hamilton et al., 2007; Linn, Baker & Betebenner, 2002).

Neben den Folgen für das Schulsystem der einzelnen Bundesstaaten, die ich hier lediglich angedeutet habe, ist eine weitere Konsequenz des NCLB-Gesetzes eine massive Zunahme der Verfügbarkeit längsschnittlicher Schülerleistungsdaten. In den USA finden daher eine Vielzahl verschiedener Modelle bzw. Methoden im Rahmen der testbasierten Evaluation von Schule und Unterricht – sowohl für bildungspolitische Zwecke, aber auch im Bildungsforschungskontext – Anwendung. Hierzu zählen Statusmodelle, stratifizierte Ranglisten, adjustierte Statusmodelle und Value-Added Modelle (vgl. Braun et al., 2010). Diese werden nachfolgend anhand ihrer zentralen Charakteristika beschrieben. Zudem werden die Gemeinsamkeiten und Unterschiede zu den bei deutschen Vergleichsarbeiten verwendeten Adjustierungsverfahren herausgearbeitet (vgl. auch Tabelle 4.4).

Statusmodelle (*status models*). In Statusmodellen wird die Testleistung der Schüler zu einem bestimmten Zeitpunkt, d. h. in einer bestimmten Klassenstufe, modelliert, ohne jedoch unterschiedliche Ausgangsvoraussetzungen des Lernens – bspw. bezüglich des SES oder der Muttersprache der Schüler – zu berücksichtigen. Nach Braun et al. (2010) liefern Statusmodelle „ … a snapshot of student performance at a point in time, which is often compared to an established target“ (Braun et al., 2010, S. 3). Dabei wird nicht definiert, was ein etabliertes Ziel bzw. Vergleichskriterium ist. In einigen Bundesstaaten der USA wird der AYP als Vergleichskriterium herangezogen (vgl. Briggs, 2008): Ein Bundesstaat legt zu Beginn des Schuljahres fest, wie viel Prozent der Schüler einer Schule *proficiency* erreichen sollen. Im Rahmen von Statusmodellen kann dann die Frage beantwortet werden, ob eine bestimmte Schule dieses Ziel, den AYP, erreicht hat (kriteriale Bezugsnorm). Als Referenz könnte ebenso ein anderes Kriterium herangezogen werden wie bspw. die durchschnittliche Testleistung aller Schüler der untersuchten Schülerpopulation (soziale Bezugsnorm). Dies entspräche dem Vorgehen in Strategie I (vgl. Abschnitt 4.1.2) im Kontext deutscher Vergleichsarbeiten.

Stratifizierte Ranglisten (*stratified league tables*). In einigen Bundesstaaten werden Ranglisten der Schulen veröffentlicht, in denen die Schulen gemäß der durchschnittlichen Testleistung ihrer Schülerschaft geordnet werden. In manchen Bundes-

staaten werden die Schulen dabei in verschiedene Strata eingeteilt, die entsprechend der sozialen Zusammensetzung der Schülerschaft (SES-Profile) gebildet werden. Dies soll die Rezipienten der Ergebnisse – dies sind v. a. Eltern auf der Suche nach einer geeigneten Schule für ihr schulpflichtiges Kind – darauf aufmerksam machen, dass die Schulen unterschiedliche Schülerpopulationen bedienen und somit nicht direkt miteinander vergleichbar sind bzw. ein solcher Vergleich *unfair* ist. Stratifizierte Ranglisten können als vereinfachte Form eines statistischen Matchings aufgefasst werden (vgl. Braun et al., 2010, S. 8). Dieses Vorgehen ist analog zu Strategie III im Kontext deutscher Vergleichsarbeiten: Auch bei Strategie III werden jeweils nur Schulen mit ähnlicher sozialer Belastung miteinander verglichen. Der wesentliche Unterschied zwischen Strategie III und stratifizierten Ranglisten liegt in der Form der Ergebnisdarstellung bzw. Rückmeldung, denn in Deutschland werden keine Ranglisten von Schulen veröffentlicht.

Adjustierte Statusmodelle (*adjusted status models*). Im Rahmen von adjustierten Statusmodellen werden die unterschiedlichen Ausgangsvoraussetzungen von Schülern berücksichtigt, indem für Kovariaten auf Schüler-, Klassen- oder/und Schulebene kontrolliert wird. Eine Möglichkeit der statistischen Modellierung adjustierter Statusmodelle ist das vom Projekt *Kompetenztest.de* bspw. im Rahmen der Hessischen Lernstanderhebungen verwendete Verfahren, d. h. Strategie IVa. Einziger Unterschied zu den nachfolgend dargestellten Value-Added Modellen ist, dass adjustierte Statusmodelle keine Kovariaten enthalten, die das Vorwissen der Schüler erfassen. Somit werden adjustierte Statusmodelle bei querschnittlichen Studiendesigns angewendet.

Value-Added Modelle (*VAM*). Der Fokus sowohl seitens der Forschung, als auch bildungspolitischer Instanzen liegt in den USA seit einigen Jahren auf Value-Added Modellen (VAM). Deren Anwendung setzt jedoch eine spezifische Datenstruktur voraus, welche gleichsam definitorisches Merkmal von VAM darstellt: die Verfügbarkeit längsschnittlicher Daten[8]. Das bedeutet, dass von jedem Schüler Testleistungsdaten zu mindestens zwei Messzeitpunkten vorliegen müssen. Meyer (1997) definiert VAM wie folgt:

[8] Wachstumsmodelle (*growth models*) und VAM sind nicht synonym, obwohl Wachstumsmodelle in der Regel auch als VAM konzeptualisiert werden können. Umgekehrt ist das nicht der Fall: Nur bestimmte VAM sind Wachstumsmodelle.

> The common characteristic of [...] value-added models [...] is that they measure school performance or the effect of school policies and inputs using a statistical regression model that includes, to the extent possible, all of the factors that contribute to growth in student achievement, in particular, student, family, and neighborhood characteristics. The key idea is to statistically isolate the contribution of schools [and/or teachers] to student achievement from all other sources of student achievement. (Meyer, 1997, S. 284)

Das gemeinsame Ziel von VAM besteht somit darin, quantitative Aussagen über die Effekte schulischer Einheiten – z. B. von Schulen und/oder Lehrern – zu treffen. Zu diesem Zweck versuchen VAM die Effekte der Lehrer, der Schule, des Schulkontextes sowie weiterer Faktoren (wie bspw. individuelle Schülercharakteristika) auf die Leistung der Schüler zu entflechten bzw. zu separieren.

Der Terminus Value-Added Modell bezieht sich nicht auf ein spezifisches statistisches Modell, sondern auf eine ganze Familie unterschiedlicher statistischer Modellierungsansätze, deren Ziel die Quantifizierung von Schul- bzw. Unterrichtseffekten ist (Braun & Wainer, 2007). Diese Modelle befinden sich in einem fortlaufenden Weiterentwicklungsprozess (vgl. z. B. Aitkin & Longford, 1986; Bryk & Weisberg, 1976; Goldstein, 1997; Hill & Rowe, 1996; Meyer, 1997; Raudenbush & Bryk, 1986; Willms & Raudenbush, 1989). Dabei gibt es in der Literatur zahlreiche Modellvarianten, wobei die Wahl des Modells von verschiedenen Faktoren abhängt. Entscheidend für die Modellwahl ist u. a. die Struktur der Daten, die Anzahl der Kohorten, die Anzahl der Messwiederholungen pro Kohorte, die Verfügbarkeit von Kovariaten, der Umgang mit fehlenden Werten etc. (vgl. Braun & Wainer, 2007, S. 875). Die resultierende Modellvielfalt führte in der VAM-Literatur zu unterschiedlichen Kategorisierungen bzw. Differenzierungen von VAM (z. B. Braun & Wainer, 2007; McCaffrey, Lockwood, Koretz, Louis & Hamilton, 2004; OECD, 2008), von denen ich im Folgenden eine näher betrachten werde.

So differenzieren McCaffrey, Lockwood, Koretz und Hamilton (2003) zwei Ansätze der Analyse längsschnittlicher Daten am Beispiel der Schätzung von Lehrereffekten: den *multivariaten Ansatz* und den *univariaten Ansatz*. Mit *multivariat* bezeichnen die Autoren solche Modelle, in denen die gesamte Datenmatrix zur Schätzung der Lehrereffekte herangezogen wird. Im Rahmen des multivariaten Ansatzes werden also die verschiedenen Outcome-Variablen, d. h. die in den verschiedenen Klassenstufen und

Fächern erhobenen Testleistungen, gemeinsam modelliert. So werden bspw. im Rahmen des *Tennessee Value-Added Assessment System* (TVAAS; Sanders & Horn, 1994; Sanders, Saxton & Horn, 1997) Testwerte aus fünf Fachbereichen und von mehreren Schülerkohorten gemeinsam modelliert. Beim univariaten Ansatz hingegen werden die Outcome-Variablen separat (jedes Outcome einzeln) bzw. sequentiell (jeder Zeitpunkt einzeln) modelliert. Hier unterscheiden McCaffrey et al. (2003) wiederum zwei unterschiedliche Modellklassen: *Gain-Score-Modelle* und *Covariate-Adjustment-Modelle* (vgl. auch Rowan, Correnti & Miller, 2002). In Gain-Score-Modellen wird die Differenz zwischen den Testleistungen von zwei aufeinanderfolgenden Schuljahren als Outcome-Variable bzw. Kriteriumsvariable betrachtet[9]. Hier wird somit explizit der Zuwachs bzw. die Veränderung der Testwerte bspw. zwischen zwei aufeinanderfolgenden Schuljahren modelliert. Will man Aussagen über die Entwicklung schulischer Leistungen zwischen zwei verschiedenen Messzeitpunkten machen, zieht dies eine zusätzliche Anforderung an die Tests nach sich: Die eingesetzten Tests müssen vertikal verlinkt sein, um die Veränderung der Testwerte auf einer gemeinsamen Skala abbilden zu können (z. B. Briggs, Weeks & Wiley, 2009). Andernfalls lassen sich positive oder negative Testwertdifferenzen nicht eindeutig auf eine Veränderung der Testwerte attribuieren, sondern können gleichfalls durch unterschiedliche Metriken, auf denen die zugrundeliegende Fähigkeit der Schüler jeweils gemessen wird, resultieren. Im Gegensatz zu Gain-Score-Modellen ist die abhängige Variable in *Covariate-Adjustment-Modellen* die Testwertvariable in einem bestimmten Schuljahr, wobei für die Kovariate Vorwissen kontrolliert wird. Die fachspezifischen Testleistungen eines Schuljahres werden als Funktion der Testwerte aus dem vorangegangenen Schuljahr, also des fachspezifischen Vorwissens – und ggf. auch weiterer Kovariaten – spezifiziert. Hier ist auch das Vorgehen des Projektes *Kompetenztest.de* zu nennen. Somit lässt sich auch Strategie IVb im Rahmen der Modellierung von VAM verorten.

[9] In der Literatur zu VAM findet sich neben Gain-Score-Modellen auch die Bezeichnung Difference-Score-Modelle. Mit Gain-Score- bzw. Difference-Score-Modellen sind jedoch nicht spezifische latente Veränderungsmodelle gemeint wie z. B. *Latent Growth Curve Models* (LGCMs; McArdle & Epstein, 1987; Meredith & Tisak, 1990) oder *True Change Models* (TCMs; McArdle & Hamagami, 2001; Raykov, 1999; Steyer, Eid & Schwenkmezger, 1997). Diese können als Spezialfälle von Gain-Score-Modellen betrachtet werden.

4.2.2 Key Stage Tests in England

Mit dem *Education Reform Act* von 1988 unter Margaret Thatcher wurde in England[10] ein nationales Curriculum eingeführt, in dem landesweit einheitliche Lerninhalte für die 5- bis 16-Jährigen zusammen mit Mindestanforderungen an Lernziele auf den unterschiedlichen Altersstufen definiert sind. Gleichzeitig wurden national verbindliche Tests (*national curriculum tests* bzw. *key stage tests*) eingeführt, die schullaufbahnbegleitend auf bestimmten Altersstufen, den sog. *Key Stages*, durchgeführt werden (z. B. H. Evans, 2008; OECD, 2008; Ray, 2006; van Ackeren, 2003a). Im nationalen Curriculum werden im Rahmen der Pflichtschulzeit vier Key Stages unterschieden: Key Stage 1 umfasst die ersten beiden Schuljahre der Grundschule. Die Testung der Schüler erfolgt am Ende des zweiten Schuljahres, bei dem die Mehrzahl der Schüler 7 Jahre alt ist. Key Stage 2 umfasst das dritte bis sechste Schuljahr. Die zugehörigen Tests werden im sechsten Schuljahr, dem Ende der Grundschulzeit, durchgeführt. Key Stage 3 umfasst die ersten drei Jahre der Sekundarstufe I, d. h. die Klassenstufen 7 bis 9, wobei die Testung wiederum am Ende dieser Phase erfolgt. Schließlich umfasst Key Stage 4 die letzten zwei Jahre der Sekundarstufe I, wobei der Großteil der Tests im elften Schuljahr durchgeführt wird. Am Ende von Klassenstufe 11 legen die Schüler das *General Certificate of Secondary Education* (GCSE) ab. Dies ist die wichtigste Abschlussprüfung für die Sekundarstufe I und entspricht dem deutschen Realschulabschluss. Mit der Evaluation der in der Folge zur Verfügung stehenden nationalen Datenbasis sind u. a. folgende Ziele verbunden (Ray, 2006, S. 5):

(1) *Öffentliche Rechenschaftslegung*: Informationen über das Leistungsniveau einzelner Schulen werden öffentlich verfügbar gemacht, indem die Testergebnisse bspw. in Form von Ranglisten veröffentlicht werden. Diese Informationen können und sollen auch Eltern im Rahmen der Schulwahl für ihre schulpflichtigen Kinder nutzen.

(2) *Schulentwicklung*: Schulen sollen die Testergebnisse im Rahmen der Selbstevaluation zur Prüfung des Erreichens der im nationalen Curriculum formulierten Lernziele verwenden und ggf. Schulentwicklungsmaßnahmen einleiten.

[10] Der Education Reform Act bezieht sich nicht auf das gesamte *Vereinigte Königreich Großbritannien und Nordirland* (*United Kingdom*; UK), sondern auf die Bildungssysteme in England, Wales und Nordirland. Die schottische Bildungsgesetzgebung ist unabhängig vom Rest des UK.

(3) *Informationen für Schulinspektionen*: Die Testergebnisse der einzelnen Schulen in den National Curriculum Tests werden zudem im Rahmen regelmäßig stattfindender Schulinspektionen, d. h. externe Evaluationsmaßnahmen zur Schulentwicklung, genutzt.

(4) *Selektion*: Die Testergebnisse werden weiterhin zur Identifikation und Auswahl von Schulen für bestimmte Förder- oder Reformmaßnahmen verwendet.

(5) *Summative Evaluation*: Die nationale Datenbasis soll weiterhin die Überprüfung der Effektivität bestimmter Schulformen und von Bildungsreformmaßnahmen ermöglichen.

Seit der Bildungsreform im Jahr 1988 wurden zum Erreichen dieser Zielstellungen verschiedene Auswertungsmodelle verwendet, wobei sich – ebenso wie in den USA – eine Entwicklung hin zu sophistizierteren statistischen Modellen nachzeichnen lässt. Diese Modelle sollen nachfolgend kurz beschrieben und wiederum zu den im Rahmen deutscher Vergleichsarbeiten verwendeten Adjustierungsverfahren in Beziehung gesetzt werden (vgl. auch Tabelle 4.4).

Ranglisten basierend auf den Rohwerten (*league tables with raw scores*). Im Jahr 1992 wurden erstmals Ranglisten von Schulen – sog. *League Tables* oder auch *Performance Tables* – veröffentlicht, die sowohl Eltern schulpflichtiger Kinder bei der Schulwahl als auch Schulen im Rahmen von Selbstevaluationsmaßnahmen unterstützen sollten. Diese Ranglisten basierten zunächst auf den Rohwerten der Testergebnisse wie bspw. den Ergebnissen aus den GCSE-Prüfungen. Ähnlich wie in Strategie I bei deutschen Vergleichsarbeiten und wie bei Statusmodellen in den USA wurde hierbei keine Adjustierung der Testwerte vorgenommen.

Value-Added Modelle (*VAM*). Die auf den unadjustierten Rohwerten basierenden Ranglisten der Schulen wurden bald als unfair angesehen, da hierbei die unterschiedliche Eingangsselektivität der Schulen nicht berücksichtigt wurde. Aufgrund der Verstetigung der National Curriculum Tests in den verschiedenen Altersstufen war seit Mitte der 1990er Jahre auch eine längsschnittliche Verknüpfung der Schulleistungsdaten aus den verschiedenen Key Stages möglich. Dies ermöglichte die Berücksichtigung der Kovariate Vorwissen im Rahmen von Value-Added Modellen. Die englischen VAM

berücksichtigten zunächst ausschließlich das Vorwissen der Schüler und waren sehr einfach gestaltet, da die resultierenden Ergebnisse für die Schulen leicht zu verstehen sein sollten (Leckie & Goldstein, 2009; Ray, 2006). Die Berechnung der Value-Added-Scores pro Schule basierte in diesen einfachen VAM auf der Median-Methode: Hierbei wird der erwartete Testwert eines Schülers mit dem beobachteten Testwert (bspw. das Outcome in Key Stage 4) verglichen. Der erwartete Testwert ist der Median der Verteilung der Testwerte aller Schüler, die den gleichen Vortestwert (bspw. hinsichtlich des Outcomes in Key Stage 3) aufweisen. Der Mittelwert der Abweichungen zwischen erwarteten und beobachteten Testwert über alle Schüler einer Schule wird dann als Maß für den *Added Value* dieser Schule, d. h. den Schuleffekt, interpretiert (vgl. Ray, 2006).

Contexualized Attainment Modelle (*CAM*). Obwohl auch in England die Verwendung von VAM – bzw. später insbesondere auch von *Contextual Value-Added Modellen* (CVA; siehe unten) als Spezialfall von VAM – zum *State of the Art* zählt, ist deren Anwendung in einigen Klassenstufen aufgrund fehlender Vorwissensmessungen nicht möglich. So wurden bspw. im Rahmen der Auswertung der Daten aus Key Stage 1 keine VAM verwendet, sondern sog. *Contexualized Attainment Modelle* (CAM; vgl. H. Evans, 2008; OECD, 2008). In diesen Modellen werden explizit keine Variablen, die das Vorwissen der Schüler erfassen, berücksichtigt. Somit können CAM auch im Rahmen querschnittlicher Studiendesigns verwendet werden. Weiterhin sind CAM synonym mit den adjustierten Statusmodellen in den USA und auch Strategie IVa bei deutschen Vergleichsarbeiten: Unterschiedliche Lernausgangsvoraussetzungen von Schülern, auf die der Lehrer bzw. die Schule keinen Einfluss hat, werden berücksichtigt, indem für diverse Kovariaten auf Schüler-, Klassen- oder/und Schulebene kontrolliert wird.

Contextual Value-Added Modelle (*CVA*). Hinsichtlich der zunächst sehr simpel gestalteten VAM wurde kritisiert, dass diese zwar einen zentralen Aspekt der differentiellen Eingangsselektivität von Schulen (das Vorwissen der Schüler) berücksichtigen, andere Faktoren – individuelle Schülereigenschaften (SES, Geschlecht etc.) oder Schulcharakteristika (Anteil von Schülern mit Migrationshintergrund etc.) – jedoch ignoriert werden. Diese Kritik und die Verfügbarkeit zusätzlicher Schülerinformationen neben

den Testleistungsdaten im Rahmen des *Pupil Level Annual School Census*[11] (PLASC) ab 2002 führte zu einer Weiterentwicklung der bis dahin verwendeten VAM. Neben dem Vorwissen wurden nun auch weitere Kovariaten – sog. Kontextfaktoren (*contextual factors*) – auf Schülerebene sowie deren Aggregate auf Schulebene berücksichtigt. Letztere umfassen zwei Variablen: (a) die durchschnittliche Testleistung der Schüler einer Schule im vorangegangenen Key Stage Test (Mittelwert) und (b) die Variabilität der Testleistungen der Schüler einer Schule im vorangegangenen Key Stage Test (Standardabweichung). Diese Modelle werden als *Contextual Value-Added Modelle* (CVA) bezeichnet[12] (Ray, 2006).

Die Besonderheit dieser Modelle besteht darin, dass diese zusätzlich zu individuellen Schülermerkmalen potenzielle *Kontext- bzw. Kompositionseffekte* berücksichtigen. Was aber sind Kontexteffekte? Und was sind Kompositionseffekte? Da diese Begriffe in unterschiedlichen Bereichen mit verschiedenen Bedeutungen konnotiert sind, sollen im Folgenden die für die vorliegende Arbeit zentralen Perspektiven – Kontexteffekte in CVA und in der Schulleistungsforschung sowie Kontexteffekte aus einer allgemeinen methodischen Perspektive – differenziert werden:

(1) *Kontexteffekte in CVA*: Im Rahmen der englischen CVA findet keine Konkretisierung dieser Begriffe statt. Hier sind mit Kontextfaktoren (*contextual factors*) ganz allgemein Faktoren gemeint, die – zusätzlich zum individuellen Lernausgangsniveau der Schüler und dem Effekt von Unterricht und Schule – Einfluss auf die Leistung bzw. den Leistungszuwachs der Schüler haben. Sämtliche Variablen, die im Rahmen des PLASC erhoben werden, sowie deren Aggregate auf Schulebene werden als Kontextfaktoren bezeichnet (vgl. H. Evans, 2008; Ray, 2006).

(2) *Kontexteffekte in der Schuleffektivitätsforschung*: In vielen Studien der Schuleffektivitätsforschung zeigt sich, dass neben den individuellen Merkmalen eines

[11] In den PLASC-Daten, die jährlich auf Schülerebene erhoben wurden, sind folgende Variablen enthalten (vgl. Ray, 2006): Geschlecht, Anspruch auf kostenloses Schulessen (als Proxy-Maß für den SES), Ethnizität, Muttersprache, Look-After-Kinder (d. h. Kinder, für die das Sorgerecht bei den Behörden liegt; bspw. Heimkinder), Zeitpunkt des Schulzuganges, Postleitzahl der Eltern. Die Informationen aus dem PLASC-Datensatz wurden mit den Testleistungsdaten der Schüler zusammengeführt und bildeten gemeinsam die *National Pupil Database* (NPD). Im Schuljahr 2006/2007 wurde PLASC durch den *School Census* ersetzt. Hierbei werden die gleichen Variablen erfasst, die Erhebung erfolgt jedoch nicht jährlich, sondern dreimal im Schuljahr.

[12] Die bei Erstellung der *School League Tables* verwendeten CVA wurden in England im Jahr 2011 wieder abgeschafft (Department for Education, 2010).

Schülers auch die Zusammensetzung der Schülerschaft – d. h. die Komposition – einer Klasse oder auch einer Schule einen Einfluss auf die Schülerleistung hat. Solche Effekte werden entsprechend als *Kompositionseffekte* bezeichnet (vgl. z. B. Baumert et al., 2006; Hattie, 2002; Rumberger & Palardy, 2005; Van Ewijk & Sleegers, 2010). Dabei deuten die bisherigen Forschungsbefunde darauf hin, dass Kompositionseffekte auf Schulebene kleiner sind als entsprechende Klassenkompositionseffekte (Van Ewijk & Sleegers, 2010; vgl. auch Maaz et al., 2011 für einen Überblick zum aktuellen Forschungsstand zu Kompositionseffekten). Man spricht jedoch nur dann von einem Kompositionseffekt, wenn das auf Klassen- oder Schulebene aggregierte Merkmal zusätzlich zu dem Individualmerkmal einen Einfluss auf die Leistung hat (Harker & Tymms, 2004; Neumann et al., 2007). Ein Beispiel soll dies verdeutlichen: Angenommen zwei Schüler mit den gleichen individuellen Lernvoraussetzungen besuchen zwei verschiedene Klassen mit unterschiedlicher Komposition der Schülerschaft. Ein Kompositionseffekt liegt dann vor, falls sich der Lernzuwachs dieser beiden Schüler – trotz vergleichbarer Unterrichtsqualität – unterscheidet. Baumert et al. (2006) unterscheiden fünf zentrale Dimensionen von Kompositionsmerkmalen: (a) die soziokulturelle Zusammensetzung, (b) die Konzentration sozialer Risikofaktoren durch belastende Familienverhältnisse, (c) die ethnisch-kulturelle Zusammensetzung, (d) das Fähigkeits- und Leistungsniveau der Schülerschaft und (e) die Konzentration lernbiographischer Belastungsfaktoren. Ein klassisches und vielfach empirisch belegtes Beispiel eines Kompositionseffektes hinsichtlich der vierten Dimension – dem Leistungsniveau der Schülerschaft – ist der sog. *Big-Fish-Little-Pond Effect* (BFLPE, vgl. z. B. Marsh, 1984, 1987; Marsh et al., 2009): Beim BFLPE ist der Effekt der individuellen Leistung der Schüler auf das akademische Selbstkonzept positiv, wohingegen der entsprechende Effekt der mittleren Leistung (aggregiert auf Klassen- oder Schulebene) negativ ist. Von Kompositionseffekten abzugrenzen sind *Institutionseffekte* unterschiedlicher Schulformen bzw. Schulsysteme. Diese beziehen sich auf „… institutionell vorgeformte Lehr-/Lernarrangements […], die ihre Verankerung in schulformspezifischen Traditionen der Didaktik und Lehrerausbildung finden“ (Baumert et al., 2006, S. 112–113).

Somit umfassen Kontexteffekte nach Baumert et al. (2006) Kompositions- und

Institutionseffekte. Diese Begrifflichkeit werde ich fortan auch im Rahmen dieser Arbeit verwenden. Eine ähnliche Differenzierung führt Willms (2008; vgl. auch Willms, 2010) ein, wobei dieser eine andere Begrifflichkeit wählt: Willms (2008) differenziert zwischen *Composition Effects* und *Contextual Effects*. Composition Effects sind Auswirkungen der demographischen Zusammensetzung einer Klasse oder Schule – wie bspw. der mittlere SES einer Klasse bzw. Schule oder die Variabilität schulischer Leistung innerhalb einer Klasse bzw. Schule – auf die Leistung einzelner Schüler. Composition Effects sind somit synonym mit Kompositionseffekten. Im Gegensatz dazu bezeichnet der Begriff Contextual Effects solche Effekte, die aufgrund der spezifischen Lehr- und Lernumgebung einer Klasse bzw. Schule zustande kommen. Damit sind z. B. Interaktionen zwischen Schülern, die Beziehung zwischen Lehrer und Schülern, das Klassenklima und die Leistungsnorm einer Schule gemeint. „Thus, it [contextual effects] comprises factors that characterize or *describe* the learning environment — its physical features, its culture and teachers' practices" (Willms, 2010, S. 1010). Contextual Effects sind somit vergleichbar mit Institutionseffekten.

(3) *Kontexteffekte aus methodischer Perspektive*: Zur Identifikation und Schätzung von Kontexteffekten – wie bspw. Effekte der Komposition der Schüler einer Klasse oder Schule – werden sog. *Contextual Models* (CM) verwendet. Dies sind spezielle Mehrebenenmodelle, die eine Variable sowohl auf Individualebene als auch auf aggregierter Ebene (z. B. Klassen- oder Schulebene) als Prädiktoren enthalten (Blalock, 1984; Raudenbush & Bryk, 2002; Snijders & Bosker, 1999). Solche Modelle werden in der Literatur auch als *Contextual Analysis Models* (z. B. Iversen, 1991) oder *Compositional Models* (z. B. Harker & Tymms, 2004) bezeichnet. Die Anwendung dieser Modelle erstreckt sich über verschiedenste Forschungsbereiche, u. a. in Medizin, Soziologie, Organisationspsychologie und Schulleistungsforschung (Iversen, 1991). Wie nun aber werden Kontexteffekte berechnet? Im Folgenden betrachten wir zu diesem Zweck eine exemplarische Analyse hierarchisch-strukturierter Daten mit zwei Ebenen: Schüler (Ebene 1), die verschiedenen Schulen (Ebene 2) angehören. Eine abhängige Variable Y, deren Werte die Mathematikleistungen der Schüler sind, soll mittels einer Ebene-1-Variable Z_{E1}, dem SES auf Schülerebene, sowie deren Aggregat auf Schulebene Z_{E2} vorhergesagt werden. Dabei sei $Z_{E2} = \bar{Z}_{\bullet j}$, d. h. die Ebene-2-Variable ist

der durchschnittlicher SES einer Schule j. Dann lautet die Modellgleichung auf Schülerebene (Ebene-1-Modell) – entsprechend der Notation[13] nach Raudenbush und Bryk (2002) – wie folgt:

$$Y_{ij} = \beta_{0j} + \beta_{1j} \cdot (Z_{ij} - \bar{Z}_{\bullet\bullet}) + r_{ij}\,, \tag{4.1}$$

wobei Y_{ij} die Mathematikleistung eines Schülers i in Schule j ist, die durch die zentrierte SES-Variable $Z_{E1} = Z_{ij} - \bar{Z}_{\bullet\bullet}$ (*grand-mean centering*; Zentrierung um den Gesamtmittelwert) vorhergesagt wird. Weiterhin sind β_{0j} das Interzept, β_{1j} der Anstieg und r_{ij} das Ebene-1-Residuum[14]. Die Modellgleichungen auf Ebene 2 eines *Random-Intercept-Modells* im Rahmen einer Kontextanalyse lauten dann folgendermaßen:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} \cdot \bar{Z}_{\bullet j} + u_{0j}\,, \tag{4.2}$$

$$\beta_{1j} = \gamma_{10}\,, \tag{4.3}$$

wobei γ die festen Effekte (*fixed effects*) und u_{0j} die zufälligen Effekte (*random effects*) sind. Die Bezeichnung als *random effects* ist hier allerdings insofern missverständlich, als u_{0j} die systematischen (und nicht zufälligen) Schulunterschiede hinsichtlich der Mathematikleistung gegeben der restlichen Prädiktoren im Modell quantifiziert. Durch Einsetzen der Ebene-2-Modellgleichungen in das Ebene-1-Modell erhält man schließlich das gemischte Modell (*linear mixed effect notation*; vgl. McCulloch & Searle, 2001):

$$\begin{aligned} Y_{ij} &= \gamma_{00} + \gamma_{10} \cdot (Z_{ij} - \bar{Z}_{\bullet\bullet}) + \gamma_{01} \cdot \bar{Z}_{\bullet j} + u_{0j} + r_{ij} \\ &= E(Y \mid Z_{E1}, Z_{E2}) + u_{0j} + r_{ij}\,. \end{aligned} \tag{4.4}$$

Ist auch der Ebene-2-Prädiktor Z_{E2} zentriert, dann ist γ_{00} die erwartete Mathematikleistung bei durchschnittlicher Ausprägung sowohl des individuellen als auch des schulspezifischen SES. γ_{10} bzw. γ_{01} quantifizieren den jeweils partiellen Zu-

[13] In Anlehnung an die in der Literatur zu Mehrebenenmodellen allgemein übliche Notationsweise verwende ich im Folgenden die Notation nach Raudenbush und Bryk (2002).

[14] Das Ebene-1-Residuum r_{ij} entspricht dem Residuum der Regression $E(Y \mid Z_{E1}, Z_{E2}, C)$ von Y auf Z_{E1}, Z_{E2} und C einer Fixed-Effects-Regression, wobei C ein Vektor mit Dummy-Variablen ist. Diese Dummy-Variablen indizieren jeweils die Schulzugehörigkeit bzw. allgemein die Zugehörigkeit zu einem Cluster (Ebene-2-Einheit).

sammenhang zwischen Mathematikleistung und dem SES auf Ebene 1 (Schülerbene) respektive auf Ebene 2 (Schulebene): γ_{10} ist die Veränderung der erwarteten Mathematikleistung bei durchschnittlicher Ausprägung der schulspezifischen SES-Variable Z_{E2}, wenn der Wert der zentrierten individuellen SES-Variable Z_{E1} um eine Einheit steigt. Dieser Parameter variiert nicht über die verschiedenen Schulen (γ_{10} ist fixiert), d. h., es wird implizit die Annahme gemacht, dass der Zusammenhang zwischen der Mathematikleistung und dem individuellen SES der Schüler in allen Schulen identisch ist. γ_{01} hingegen ist die Veränderung der erwarteten Mathematikleistung bei durchschnittlicher Ausprägung der individuellen SES-Variable Z_{E1}, wenn der Wert der auf Schulebene aggregierten SES-Variable Z_{E2} um eine Einheit steigt. Dieser Parameter ist ein Schätzer für den Kontexteffekt. Ein Kontexteffekt liegt dann vor, wenn $E(Y|Z_{E1}, Z_{E2}) \neq E(Y|Z_{E1})$, d. h., wenn gilt:

$$\gamma_{01} \neq 0\,. \tag{4.5}$$

Hat also der durchschnittliche SES einer Schule (Z_{E2}) zusätzlich zum individuellen SES der Schüler (Z_{E1}) einen Effekt auf die Mathematikleistung Y, so liegt ein Kontexteffekt vor.

In Mehrebenenmodellen ist die Wahl der Lokation des Prädiktors, d. h. die Zentrierung, entscheidend für die Bedeutung und Interpretation der Regressionskoeffizienten (vgl. z. B. Raudenbush & Bryk, 2002). Üblich sind (a) die bereits erwähnte Zentrierung um den Gesamtmittelwert sowie (b) die Zentrierung um den Gruppenmittelwert des Ebene-1-Prädiktors (*group-mean centering*). Auch in Contextual Models (CM) ist es möglich und üblich, den Ebene-1-Prädiktor um den Gruppenmittelwert zu zentrieren (Enders & Tofighi, 2007; Lüdtke et al., 2008). Das resultierende Modell ist – im Falle des Random-Intercept-Modells – mathematisch äquivalent zu dem in Gleichung 4.4 dargestellten Modell (Kreft, de Leeuw & Aiken, 1995). Jedoch verändert sich in der Folge die Bedeutung der Regressionskoeffizienten: Der Kontexteffekt ergibt sich dann indirekt als Differenz zweier Modellparameter. Dies ist ausführlich im Anhang C dargestellt.

Hinsichtlich der kausalen Interpretierbarkeit der Parameter als (ursächliche) Kontexteffekte ist jedoch – ebenso wie bei den im Rahmen dieser Arbeit fokussierten fairen Vergleichen – Vorsicht geboten: Kontexteffekte können nicht ohne weitere Annahmen kausal interpretiert werden (vgl. Blalock, 1984; Marsh et al., 2009).

In vielen Studien sind Kontextvariablen (d. h. Ebene-2-Prädiktoren) Aggregate der Individualmerkmale, also bspw. gruppenspezifische Mittelwerte der manifesten Variablen auf Ebene 1. Da dies im Falle unreliabler Messungen der Kontextvariablen zu sog. *phantom compositional effects* führen kann (z. B. Harker & Tymms, 2004), werden im Rahmen weiterentwickelter Ansätze latente Kontextvariablen modelliert, um dem Messfehlerproblem zu begegnen (Lüdtke et al., 2008; Marsh et al., 2009).

4.2.3 USA, England und Deutschland im Vergleich

Tabelle 4.4 gibt einen zusammenfassenden Überblick der verwendeten Adjustierungsmodelle in den drei Ländern USA, England und Deutschland. In den Spalten 2 bzw. 3 sind die Modellbezeichnungen aus den USA respektive England aufgelistet. Die Tabelle weist zudem in Spalte 4 die Beziehung zu den in deutschen Vergleichsarbeiten verwendeten Adjustierungsverfahren aus.

In der ersten Spalte sind die zentralen Modellcharakteristika bzw. die jeweils verwendeten Kovariaten aufgelistet, wobei die Modelle von (1) bis (6) eine wachsende Komplexität und auch zunehmende Fairness aufweisen. Zusätzlich sind die Modelle verschiedenen Modelltypen zugeordnet, die auf der Klassifikation von Value-Added Modellen nach Timmermans et al. (2011) basiert. Die Autoren unterscheiden in ihrer Klassifikation verschiedene Typen von VAM (Type 0, Type AA, Type A und Type B), die jeweils zu unterschiedlichen Interpretationen der resultierenden Effektschätzungen führen: Das Effektmaß aus VAM vom Type 0 ist demnach lediglich als Differenz zwischen einer konkreten Schule und der durchschnittlichen Schule bezüglich der mittleren Leistung der Schüler zu interpretieren. Dahingegen liefern VAM vom Type AA ein Effektmaß, das als Differenz der durchschnittlichen Schülerleistung zwischen einer konkreten Schule und der durchschnittlichen Schule bei gegebenem Vorwissensniveau der Schüler zu verstehen ist. Die Differenzierung in Type A und Type B ist synonym mit der Klassifikation nach Raudenbush und Willms (1995) in Type-A-Effekte und Type-B-Effekte.

Type-A-Effekte vs. Type-B-Effekte. Raudenbush und Willms (1995) sowie Willms und Raudenbush (1989) argumentieren, dass sich die Schulleistung eines Schülers (das Outcome) als Funktion von vier Faktoren darstellen lässt: individuelle Schülermerk-

Tabelle 4.4: Adjustierungsmodelle: USA, England und Deutschland im Vergleich

Modellcharakteristika/ Kovariaten	Länder		
	USA	England	Deutschland
(1) Keine Adjustierung (Type 0)[a]	Statusmodelle	Ranglisten basierend auf Rohwerten	Strategie I
(2) Stratifizierung basierend auf SES-Maßen	Stratifizierte Ranglisten	—	Strategie III[b]
(3) Schüler- und Kompositions-merkmale; ohne Vorwissen	Adjustierte Statusmodelle	CAM	Strategie IVa
(4) Vorwissen (Type AA)[a]	VAM	VAM	—
(5) Vorwissen; Schülermerkmale (Type A)[a]	VAM[c]	—	Strategie IVb
(6) Vorwissen; Schüler- und Kompositions-merkmale (Type B)[a]	VAM[c]	CVA[d]	—

Anmerkungen. CAM = Contextualized Attainment Model, VAM = Value-Added Model, CVA = Contextual Value-Added Model.

[a] Die Differenzierung in Type 0, Type AA, Type A und Type B basiert auf der Klassifikation von Value-Added Modellen nach Timmermans, Doolaard und de Wolf (2011).

[b] Auch Strategie II ist vergleichbar mit dem Vorgehen bei stratifizierten Ranglisten, da es sich gleichfalls um eine vereinfachte Form eines statistischen Matchings handelt. Lediglich die Art der Kovariaten, hinsichtlich derer *gematcht* wird, unterscheidet sich, denn bei Strategie II werden keine Maße des SES verwendet.

[c] In den USA wird keine weitere begriffliche Differenzierung hinsichtlich VAM, die zusätzlich Kovariaten auf Schülerebene oder/und Kompositionsmerkmale auf Klassen- oder Schulebene modellieren, vorgenommen. Gleichwohl werden auch diese Modelle angewandt.

[d] Die bei Erstellung der *School League Tables* verwendeten CVA wurden in England im Jahr 2011 wieder abgeschafft (Department for Education, 2010).

male, Messfehler, Schulkontext und Schulpraxis. Der Begriff Schulpraxis ist hier sehr weit gefasst und umfasst die administrative Leitung einer Schule, curriculare Inhalte, Ressourcenverwendung und den Unterricht in den Klassen. Im Gegensatz dazu umfasst der Schulkontext Faktoren auf Schulebene, auf die die Schulleitung und Lehrer keinen Einfluss haben – wie bspw. die soziale Komposition der Schülerschaft. Basierend auf diesem Verständnis des Zustandekommens schulischen Outputs, d. h. der Leistung der Schüler, definieren die Autoren zwei Arten von Schuleffekte, die in einem Educational-Accountability-System von Interesse sind: (a) den *Type-A-Effekt* und (b) den *Type-B-Effekt* (vgl. auch Raudenbush, 2004). Beide Effekte sind definiert als Differenz zwischen der tatsächlichen Leistung der Schüler einer bestimmten Schule und der erwarteten Leistung dieser Schüler, wären diese in einem anderen schulischen Setting. Unterschiede hinsichtlich Interpretation der Schuleffekte ergeben sich durch die Wahl der Referenz bzw. des Vergleichssettings:

In Modellen zur Schätzung von Type-A-Effekten werden individuelle Schülereigenschaften (z. B. individuelles Vorwissen, SES, Geschlecht) als Kovariaten berücksichtigt. Der Type-A-Effekt ist dann die Differenz zwischen der tatsächlichen Leistung der Schüler einer bestimmten Schule und der erwarteten Leistung vergleichbarer Schüler einer hinsichtlich des Schulkontextes und der Schulpraxis durchschnittlichen Schule. Dieser Effekt ist v. a. für Eltern relevant, die im Rahmen freier Schulwahl eine Schule für ihre Kinder wählen: Für diese Entscheidung ist nicht nur der Effekt der Schulpraxis entscheidend, sondern vielmehr das Zusammenspiel von Schulpraxis und Schulkontext auf die Effektivität einer Schule.

Im Gegensatz dazu werden in Modelle zur Schätzung von Type-B-Effekten zusätzlich zu den genannten individuellen Schülermerkmalen Kompositionsmerkmale der Schule (z. B. durchschnittliches Vorwissen der Schüler einer Schule) als Kovariaten aufgenommen. Somit wird auch der Schulkontext einer Schule bei der Adjustierung berücksichtigt, mit dem Ziel, den Effekt der Schulpraxis zu separieren und zu quantifizieren. Der Type-B-Effekt ist die Differenz zwischen der tatsächlichen Leistung der Schüler einer bestimmten Schule und der erwarteten Leistung vergleichbarer Schüler an einer Schule mit gleichem Schulkontext, die hinsichtlich der Schulpraxis durchschnittlich ist. Wollen also der Schulleiter und die Lehrer einer Schule eine Evaluation der eigenen schulischen Arbeit vornehmen, ist der Type-B-Effekt relevant. So kann bspw. der Type-A-Effekt einer Schule aufgrund einer vorteilhaften sozialen Komposition der Schülerschaft sehr positiv ausfallen, der Type-B-Effekt jedoch negativ sein.

Das bedeutet, dass diese Schule schlechtere Ergebnisse hinsichtlich der durchschnittlichen Leistung ihrer Schülerschaft beim Vergleich mit Schulen vergleichbarer sozialer Zusammensetzung zeigt. Ein solches Ergebnis kann wiederum Ausgangspunkt schulischer Qualitätsentwicklungsmaßnahmen sein. Daher zielt man auch im Rahmen eines Educational-Accountability-Systems auf die Schätzung des Type-B-Effekts.

Die Differenzierung in Type-A-Effekte und Type-B-Effekte ist in der Praxis der Schulleistungsforschung jedoch nicht derart trennscharf. So ist es an manchen Schulen durchaus möglich, dass die Schulleitung bzw. Lehrer gleichfalls einen Einfluss auf die soziale Zusammensetzung ihrer Schülerschaft haben. Die Separierung von Faktoren, auf die Schule und Unterricht Einfluss haben vs. Faktoren, die sich der Kontrolle der Schule entziehen, ist demnach keine triviale Frage und kann nicht universell beantwortet werden. Des Weiteren gilt auch hier: Um kausal interpretierbare Effektschätzungen von Type-A-Effekten respektive Type-B-Effekten zu erhalten, müssen sämtliche relevante Kovariaten sowie das korrekte statistische Modell gewählt werden. Nach (Raudenbush, 2004) lässt sich vor diesem Hintergrund lediglich die Schätzung des Type-A-Effekts rechtfertigen. Im Gegensatz dazu sind Chancen zur Schätzung des kausalen Type-B-Effekts „… dim at best“ (Raudenbush, 2004, S. 123).

Ein Zwischenfazit. Unabhängig von den *Tücken* kausaler Effektschätzung macht die konzeptuelle Unterscheidung dieser beiden Arten von Effekten (Type-A-Effekte und Type-B-Effekte) Folgendes deutlich: Ein zentrales Kriterium für die Wahl des geeigneten Adjustierungsverfahrens ist das Ziel bzw. die übergeordnete Fragestellung, zu deren Beantwortung Klassen- und/oder Schuleffekte geschätzt werden (vgl. auch Briggs, 2008): Type-A-Effekte beantworten die Frage der Schulwahl für Eltern. Hierfür sind klassische VAM geeignet, die zur Adjustierung das Vorwissen der Schüler sowie individuelle Schülermerkmale berücksichtigen, auf die Schule keinen Einfluss hat. Für die Evaluation schulischer Arbeit – also des Ertrags von Unterricht und/oder Schule – sollten hingegen Type-B-Effekte betrachtet werden. Zu diesem Zweck werden – bspw. im Rahmen von CVA – zusätzlich Kontexteffekte wie bspw. die leistungsmäßige Komposition einer Klasse und/oder Schule berücksichtigt. Dementsprechend kritisch merkten Leckie und Goldstein (2009) im Kontext der englischen CVA an, dass „… the contextual value-added estimates do include school compositional effects and are therefore not appropriate for choice purposes. It is thus somewhat ironic that they have been promoted by government as improving choice“ (Leckie & Goldstein, 2009, S. 838).

4.3 Zusammenfassung

Landesweite Vergleichsarbeiten sind seit dem Jahr 2006 Teil der KMK-Gesamtstrategie zum Bildungsmonitoring des deutschen Bildungssystems. Die Ergebnisse dieser standardisierten Testverfahren sollen – neben verschiedenen anderen Zielen – Aussagen über Unterrichtseffekte ermöglichen und somit Ausgangspunkt für Unterrichtsentwicklung sein können. Dazu werden die Testergebnisse einer Klasse mit den Ergebnissen anderer Klassen verglichen (soziale Bezugsnorm). Um faire(re) Vergleiche zu ermöglichen, müssen dabei auch außerschulische Einflussgrößen des Lernens, die ebenfalls die Schülerleistung beeinflussen (Kovariaten), in der Analyse der Testergebnisse berücksichtigt werden. Obwohl bezüglich der Ziele mittlerweile ein weitestgehend einheitlicher Rahmen vorliegt, zeigt sich bei näherer Betrachtung der Vergleichsarbeiten über die einzelnen Bundesländer hinweg ein sehr heterogenes Bild. Unterschiede bestehen nicht nur in Bezug auf die Bezeichnung der Vergleichsarbeiten, sondern auch im Hinblick auf die Testdurchführung sowie insbesondere hinsichtlich der Datenanalyse und Rückmeldung der Ergebnisse.

Der Fokus dieser Arbeit liegt auf der statistischen Datenanalyse, d. h. der Analyse der Testergebnisse als zentrale Schnittstelle zwischen der Messung (Erfassung von Schülerleistungen und außerschulischen Einflussgrößen des Lernens) und der Ergebnisrezeption. Insbesondere für die Analysestrategien im Rahmen der Vergleichsarbeiten zeigt sich über die Bundesländer kein einheitliches Bild, wobei sich im Wesentlichen vier verschiedene Vorgehensweisen differenzieren lassen. So werden teilweise noch unadjustierte Vergleichswerte zurückgemeldet. Werden Adjustierungen durchgeführt, dann beziehen sich diese auf die Berechnung des Vergleichswertes. Die Differenz des beobachteten Klassenmittelwertes vom jeweils berechneten adjustierten Vergleichswert soll dann als Maß der Effektivität des Unterrichts interpretiert werden können. Aus methodischer Sicht ist dabei der Fairness-Aspekt zentral: Nur wenn alle relevanten Kovariaten in der Analyse adäquat, d. h. mit Hilfe des richtigen statistischen Modells, berücksichtigt werden, können die berechneten Differenzwerte als ursächliche Effekte des Unterrichts interpretiert werden. In der Praxis empirischer Bildungsforschung spielen jedoch stets auch Praktikabilitätsaspekte wie Testökonomie oder Modellkomplexität eine nicht zu vernachlässigende Rolle. So können bspw. aus testökonomischen Gründen nicht alle relevanten Kovariaten erfasst werden. Auch das verwendete statistische Modell muss einem Sparsamkeitskriterium genügen, nicht zuletzt um die Transparenz und Kommu-

nizierbarkeit der berechneten Effektmaße zu gewährleisten.

Ein Vergleich mit anderen Educational-Accountability-Systemen macht zudem deutlich, dass sich die Adjustierungsstrategien im Rahmen deutscher Vergleichsarbeiten den Vorgehensweisen in den beiden Ländern USA und England zuordnen lassen. Im Gegensatz zu Deutschland ist die Evaluation der Testergebnisse insbesondere in den USA mit stärkeren Konsequenzen, wie bspw. Sanktionen für Schulen oder Lehrer, verknüpft. Dies wird auch als *High-Stakes Assessment System* bezeichnet. Die Vorgehensweise in Rahmen deutscher Vergleichsarbeiten lassen sich hingegen einem sog. *Low-Stakes Assessment System* zuordnen, denn öffentliche Rankings von Schulen und Lehrern sowie zentrale Sanktionierungen existieren – erfreulicherweise – nicht. Der Vergleich macht jedoch in erster Linie auf potenzielle Erweiterungsmöglichkeiten hinsichtlich der Adjustierung von Testergebnissen aufmerksam. Diese sind Ausgangspunkt für die im folgenden Kapitel darzustellenden Fragestellungen sowie Hypothesen.

The question of the question is always the biggest question.

THOMAS D. COOK (2010)

5 Fragestellungen

Im nachfolgenden Kapitel werden die zentralen Befunde des theoretischen Teils dieser Arbeit zusammengefasst. Die Problematik fairer(er) Vergleiche im Kontext von Vergleichsarbeiten wird sodann anhand zweier zentraler Facetten präzisiert (Abschnitt 5.2). Des Weiteren werden drei methodische Zugänge differenziert, mittels derer die im Rahmen dieser Arbeit betrachteten Teilfragen fairer(er) Vergleiche analysiert werden können (Abschnitt 5.3). Diese drei methodischen Zugänge werden zudem mittels aktueller Forschungsbefunde illustriert. Darauf aufbauend werden die zentralen Forschungsfragen dieser Arbeit dargestellt (Abschnitt 5.4). Vor dem Hintergrund dieser Fragestellungen wird der methodische Zugang im Rahmen der vorliegenden Arbeit festgelegt. Abschließend werden die Hypothesen spezifiziert, die im empirischen Teil dieser Arbeit einer Überprüfung an Beobachtungsdaten unterzogen werden.

5.1 Faire Vergleiche als kausale Unterrichtseffekte

In den vorangegangenen Kapiteln wurde die Heterogenität von Vergleichsarbeiten sowohl bezüglich der Zielstellungen (vgl. Kapitel 2) als auch der Durchführungspraxis (v. a. hinsichtlich der Datenanalyse und Rückmeldung von Ergebnissen aus Vergleichsarbeiten; vgl. Kapitel 4) herausgearbeitet. Im Rahmen der vorliegenden Arbeit fokussiere ich insbesondere das Ziel, mittels Vergleichsarbeiten zu fairen, d. h. kausal interpretierbaren Vergleichen zu gelangen: Ein zentrales Ziel von Vergleichsarbeiten ist – explizit oder implizit – die Quantifizierung von Unterrichtseffekten auf die Schülerleistung.

Ausgehend von dieser Zielstellung, kausale Unterrichtseffekte mittels Ergebnissen aus Vergleichsarbeiten zu quantifizieren, stellt sich die Frage, wie man zu fairen Vergleichen kommen kann. In Kapitel 3 habe ich daher die zentralen Konzepte und Annahmen der allgemeinen stochastischen Theorie kausaler Effekte (Steyer et al., 2011)

eingeführt. Weiterhin habe ich herausgearbeitet, dass im Kontext von Vergleichsarbeiten insbesondere der *ACE on the treated* von Interesse ist (vgl. Abschnitt 3.3.2). Analytisch, d. h. mathematisch, wurde gezeigt, dass der *ACE on the treated* unter Gültigkeit bestimmter Annahmen durch empirisch schätzbare Größen identifiziert werden kann (vgl. Kapitel 3). Dieser Befund bezieht sich auf die kausale Inferenz, d. h. die Inferenz von Populationsparametern auf kausale Effekte (vgl. Abschnitt 3.1 und Abbildung 3.1). Darüber hinaus wurde gezeigt, dass sich das tatsächlich praktizierte Vorgehen zur Berechnung fairer Vergleiche aus kausaltheoretischer Perspektive rechtfertigen bzw. deduzieren lässt, jedoch mit einer Reihe sehr starker Annahmen verknüpft ist (vgl. Kapitel 3 und Fiege, 2007): Wie in Kapitel 3 (Abschnitt 3.5) ausführlich dargestellt ist der in Vergleichsarbeiten jeweils berechnete adjustierte Referenzwert $E[\, E(Y \,|\, \boldsymbol{Z}) \,|\, X{=}x \,]$ zunächst lediglich der für eine Klasse bzw. Schule zu erwartende Wert adjustiert für die konkreten Variablen $\boldsymbol{Z}$ im Adjustierungsmodell. Werden in einem nächsten Schritt zusätzliche Variablen $\boldsymbol{Z}^+$ im Adjustierungsmodell berücksichtigt, so kann sich ein anderer Referenzwert ergeben. Der nun resultierende Vergleichswert $E[\, E(Y \mid \boldsymbol{Z}, \boldsymbol{Z}^+) \mid X{=}x \,]$ ist der für eine Klasse bzw. Schule zu erwartende Wert adjustiert für $\boldsymbol{Z}$ *und* $\boldsymbol{Z}^+$. Dieser kann sich somit von dem zuerst berechneten adjustierten Vergleichswert $E[\, E(Y \mid \boldsymbol{Z}) \mid X{=}x \,]$ unterscheiden. Zudem werden sich beide Werte von dem kausalen Vergleichs- bzw. Referenzwert unterscheiden, sofern nicht die Annahme der bedingten Unverfälschtheit erfüllt ist *und* Gleichung 3.31 in Kapitel 3 gilt.

Die Plausibilität dieser Annahmen in konkreten empirischen Anwendungen – hier also im Kontext von Schulleistungsuntersuchungen bzw. Vergleichsarbeiten – ist fraglich. Dies wird gleichfalls in der Literatur kritisch diskutiert (z. B. Briggs, 2008; Raudenbush, 2004; Rubin, Stuart & Zanutto, 2004). Zudem lassen sich im Rahmen des in Kapitel 3 gewählten analytischen Zuganges, bei dem ein Vergleich der theoretischen mit den empirisch schätzbaren Größen betrachtet wird, keine Aussagen über die Richtigkeit bzw. Angemessenheit dieser Annahmen ableiten. Des Weiteren lassen sich diese Annahmen in konkreten empirischen Anwendungen – also gleichfalls in den im Rahmen dieser Arbeit betrachteten Schulleistungsuntersuchungen (bzw. allgemein in Beobachtungsstudien) – weder herstellen noch verifizieren, sondern allenfalls falsifizieren. Lediglich (theoretische) Plausibilitätsüberlegungen auf Basis der konkreten Gegebenheiten im Anwendungsfall sind möglich.

Aus diesen Gründen können die berechneten adjustierten Effektmaße stets nur eine Annäherung an Unterrichtseffekte – und in diesem Sinne lediglich *fairere* Vergleiche –

darstellen. Die berechneten Maße können nicht als ursächliche – also kausale – Effekte des Unterrichts verstanden werden, sondern sollten als *deskriptive Maße* interpretiert werden (vgl. Briggs, 2008; Rubin et al., 2004).

In der Zusammenschau der Argumente, die für bzw. gegen die Plausibilität dieser Annahmen bei Vergleichsarbeiten sprechen, komme ich zu dem Schluss, dass die Schätzung kausaler Unterrichtseffekte in diesem Bereich theoretisch möglich ist. Aus praktischer Sicht ist dies jedoch unrealistisch: Allein die *erschöpfende* Erfassung aller relevanter Kovariaten unterliegt im Kontext von Vergleichsarbeiten deutlichen Grenzen, u. a. hinsichtlich der begrenzten kognitiven, zeitlichen und finanziellen Ressourcen der Schüler, Lehrer bzw. des jeweiligen landesspezifischen Bildungssystems sowie datenschutzrechtlicher Limitationen. Mittels der praktizierten Adjustierungen lässt sich bestenfalls eine Näherung an die intendierten kausalen Effekte – und damit lediglich *fairere Vergleiche* – erreichen. High-Stakes Assessment mit faireren Vergleichen als alleinigem Gütekriterium ist vor diesem Hintergrund m. E. nicht haltbar. Es ist daher als positiv zu beurteilen, dass die Ergebnisse aus Vergleichsarbeiten nicht für die Sanktionierung von Lehrkräften (bspw. Gehaltskürzungen) oder Schulen (bspw. Schulschließungen) verwendet werden.

Trotz der erwähnten Einschränkungen bergen solche adjustierten, faireren Vergleichswerte ein großes Potenzial, um als Informationsbasis für die beteiligten Lehrkräfte zu fungieren. Die so berechneten Vergleichswerte können Ausgangspunkt kollegialer Diskussionen sein sowie Impulse für Unterrichtsentwicklungsmaßnahmen geben. Eine zentrale Voraussetzung dafür ist, dass die Möglichkeiten, aber auch die Grenzen der Interpretation transparent dargestellt werden (vgl. Maier, 2008).

5.2 Zwei zentrale Facetten fairer(er) Vergleiche

Wie kann die Kausalitätstheorie bei der Entwicklung und Bewertung einer geeigneten Adjustierungsstrategie dennoch weiterhin nützlich sein? Betrachtet man die Problematik fairer Vergleiche vor dem Hintergrund einer allgemeinen Theorie kausaler Effekte, so lässt sich diese Fragestellung durch die folgenden zwei Teilfragen präzisieren:

(1) Welche Kovariaten müssen im statistischen Modell zur Schätzung von Unterrichtseffekten berücksichtigt werden? (Kovariatenselektion)

(2) Welches ist das richtige statistische Modell zur Schätzung von Unterrichtseffekten? (Modellselektion)

Die Missspezifikation eines Modells kann demnach im Wesentlichen zwei Ursachen haben: (a) Wichtige Variablen wurden nicht in das Modell aufgenommen (*omitted variables*) oder (b) die modellierte entspricht nicht der wahren funktionalen Form der Abhängigkeit zwischen den im Modell enthaltenen Zufallsvariablen. Wichtig ist, dass diese beiden Fragen nicht unabhängig voneinander beantwortet werden können, sondern stets gemeinsam betrachtet werden müssen. So ist in einer konkreten empirischen Anwendung die Wahl eines adäquaten statistischen Modells stets abhängig von den stochastischen Abhängigkeiten bzw. den gemeinsamen Verteilungen der darin enthaltenen Zufallsvariablen. Darüber hinaus entscheidet die Variablenselektion gleichfalls darüber, welches statistische Modell verwendet werden kann. So ist bspw. die Variable Vorwissen (bzw. Prätest) definitorischer Bestandteil eines Value-Added Modells (VAM). Zum Zwecke der Präzisierung der Fragestellung dieser Arbeit und der Erarbeitung eines konkreten methodischen Zuganges ist die Differenzierung in die beiden Facetten bzw. Teilfragen Kovariaten- und Modellselektion jedoch von zentraler Bedeutung.

5.3 Aggregation bisheriger Befunde aus der Perspektive verschiedener methodischer Zugänge

Betrachtet man die verschiedenen Strategien der Datenanalyse und Rückmeldung von Ergebnissen aus Vergleichsarbeiten (vgl. Kapitel 4), so fällt die Heterogenität der Adjustierungsverfahren bezüglich der beiden Facetten – sowohl der Kovariaten- als auch der Modellselektion – auf. Auch in der amerikanischen und europäischen Literatur zu Schulleistungsstudien gibt es bisher keinen allgemeinen Konsens über die Wahl der Variablen und des statistischen Modells (vgl. z. B. Braun & Wainer, 2007). Hinsichtlich der Adjustierungsverfahren im Kontext deutscher Vergleichsarbeiten ist die Befundlage sogar äußerst spärlich: Die verschiedenen Adjustierungsverfahren bzw. -modelle waren bisher kaum Gegenstand derartig vergleichender Analysen. Jedoch finden sich in der Literatur zu statistischen Modellen im Kontext der Schulleistungsforschung (insbesondere in der Literatur zu Value-Added Modellen) verschiedene methodische Zugänge,

mittels derer man diese beiden Fragen zu beantworten sucht. Diese methodischen Zugänge lassen zum Teil unterschiedliche Implikationen zu. Im Einzelnen lassen sich die folgenden Ansätze bzw. methodischen Zugänge unterscheiden (vgl. z. B. McCaffrey et al., 2003; Braun & Wainer, 2007): (a) der analytische Zugang, (b) der simulationsbasierte Zugang und (c) der empirische Zugang. Zentral ist der ergänzende Charakter dieser drei methodischen Zugänge: Zwar lassen sich bestimmte Fragestellungen – wie bspw. die Äquivalenz zweier statistischer Modelle – aus allen drei Perspektiven analysieren. Wiederum andere Fragestellungen – wie bspw. das Ausmaß des Einflusses von zusätzlichen Variablen im Modell auf die konkreten Effektschätzungen – lassen sich ausschließlich mittels des empirischen Zuganges beantworten[1].

Die drei methodischen Zugänge – analytisch, simulationsbasiert und empirisch – werden nachfolgend beschrieben sowie hinsichtlich der Möglichkeiten und Grenzen ihrer Aussagekraft bzw. der Generalisierbarkeit ihrer Ergebnisse kritisch diskutiert. Vor dem Hintergrund des jeweiligen methodischen Zuganges werden Beispiele von Schulleistungsstudien mit Befunden zu den beiden Teilfragen – Kovariaten- und Modellselektion – berichtet. Tabelle 5.1 zeigt eine Übersicht der dafür ausgewählten Studien und ordnet diese hinsichtlich Forschungsfrage und methodischem Zugang. An dieser Stelle sei darauf hingewiesen, dass die Übersicht keinesfalls erschöpfend ist, sondern lediglich exemplarisch ausgewählte Befunde aufgreift, wobei der Fokus auf Untersuchungen zu Value-Added Modellen liegt. Auf der Basis dieser Befunde werden schließlich in Abschnitt 5.4 die zentralen Forschungsfragen sowie die Wahl des empirischen Zuganges im Rahmen dieser Arbeit herausgearbeitet.

5.3.1 Der analytische Zugang

Die analytische Herangehensweise zeichnet sich dadurch aus, dass mittels mathematischer Operationen und Gesetzmäßigkeiten (z. B. Rechenregeln für Erwartungswerte etc.), d. h. mittels mathematischer Beweise, die Berechnung eines statistischen Parameters – bspw. des Standardfehlers eines statistischen Kennwertes – formal-logisch hergeleitet wird. Auch die Äquivalenz verschiedener statistischer Parameter oder von statistischen Modellen lässt sich auf diese Weise zeigen.

[1] Eine derartige Fragestellung lässt sich gleichfalls mittels des simulationsbasierten oder auch analytischen Zuganges betrachten. Jedoch sind hier Annahmen bspw. über die Verteilungen von Zufallsvariablen nötig, die wiederum in der Regel aus empirischen Beobachtungen bzw. Studien resultieren.

Tabelle 5.1: Wissenschaftliche Studien zu den beiden Facetten fairer(er) Vergleiche vor dem Hintergrund verschiedener methodischer Zugänge

Methodischer Zugang	Facetten fairer(er) Vergleiche	
	Kovariatenselektion	Modellselektion
(1) analytisch	McCaffrey et al. (2004)	McCaffrey et al. (2004) Fiege (2007)
(2) simulationsbasiert	McCaffrey et al. (2004)	McCaffrey et al. (2004)
(3) empirisch	Ballou et al. (2004) Tekwe et al. (2004) Hedges & Hedberg (2007) Leckie & Goldstein(2009) Benton et al. (2003) Timmermans et al. (2011) Briggs & Domingue (2011)	Tekwe et al. (2004)

Anmerkungen. Aufgeführt sind Beispiele für Studien, in denen die Kovariaten- oder Modellselektion im Kontext von Schulleistungsuntersuchungen untersucht wurde. Mehrfachnennungen indizieren, dass entweder beide Teilfragen fairer Vergleiche betrachtet oder/und verschiedene methodische Zugänge gewählt wurden.

So wählen McCaffrey et al. (2004) den analytischen Zugang, um die Äquivalenz verschiedener Value-Added Modelle zu untersuchen. Die Autoren schlagen in ihrem Artikel ein allgemeines Value-Added Modell vor: das sog. *variable persistence model*. Sie zeigen analytisch, dass die bis dahin am häufigsten verwendeten Value-Added Modelle Spezialfälle dieses Modells darstellen. Die Autoren zeigen jedoch weiterhin, dass dies nicht automatisch die Überlegenheit des allgemeinen Modells impliziert. Entscheidend ist die gemeinsame Verteilung der relevanten Kovariaten. Zudem bestätigen sie diesen Befund mit Ergebnissen einer zusätzlichen Simulationsstudie. Braun und Wainer (2007) resümieren das Fazit dieser Untersuchung folgendermaßen: „The overall conclusion is that no one model dominates the others under all plausible circumstances. The optimal strategy depends on how the covariates are distributed across classes, schools and groups of schools“ (S. 884).

Bezogen auf Adjustierungsverfahren im Kontext von Vergleichsarbeiten ist die Befundlage aus Studien mit analytischem Zugang übersichtlich: Fiege (2007) zeigt, dass

das in der nationalen Erweiterung von PISA 2000 (PISA-E 2000) verwendete Adjustierungsverfahren ein Spezialfall des Adjustierungsverfahrens im Projekt *Kompetenztest.de* ist: Bei der Adjustierung im Projekt *Kompetenztest.de* werden keine Annahmen über die funktionale Form der Abhängigkeit der Testwertvariablen (d. h. der Outcome-Variablen Y) von dem Kovariatenvektor $\boldsymbol{Z}$ gemacht, die sich in empirischen Anwendungen als falsch erweisen können. Dahingegen wurde im Rahmen von PISA-E 2000 ein restriktiveres Modell gewählt. Hierbei wurde ein linearer Zusammenhang zwischen der Outcome-Variablen Y und dem Kovariatenvektor $\boldsymbol{Z}$ modelliert, wobei potenzielle Interaktionen zwischen den Kovariaten nicht berücksichtigt wurden. Dem Vorteil eines sparsameren Modells (Parsimonitätsprinzip), steht der Nachteil gegenüber, einen wahren nonlinearen Zusammenhang nicht abbilden zu können und somit potenziell falsche Parameterschätzungen zu erhalten. Über die Plausibilität der Linearitätsannahme lässt sich jedoch im Rahmen des analytischen Zuganges keine Aussage treffen.

Beide Studien – sowohl von McCaffrey et al. (2004) als auch von Fiege (2007) – machen gleichfalls auch die Grenze der analytischen Methode deutlich: So kann man analytisch zwar zeigen, dass zwei Modelle nicht äquivalent sind. Ohne Kenntnis der wahren Zusammenhänge bzw. Abhängigkeiten der im Modell enthaltenen Variablen lassen sich die Auswirkungen auf die Parameterschätzung jedoch nicht quantifizieren. Wie McCaffrey et al. (2004) zeigen, können in diesem Fall Simulationsstudien zusätzliche Informationen und Erkenntnisse liefern.

5.3.2 Der simulationsbasierte Zugang

In der Statistik lassen sich im Wesentlichen zwei Anwendungsvarianten von Simulationsstudien unterscheiden (vgl. z. B. Mooney, 1997; Rubinstein, 1981): Einerseits werden diese zur Erzeugung der Kennwerteverteilung einer Teststatistik angewendet, um den Standardfehler eines statistischen Parameters empirisch zu bestimmen. Ein simulationsbasierter Zugang ermöglicht andererseits eine Abschätzung des Fehlers (bzw. des sog. *Bias*), den man macht, wenn man trotz Verletzungen von Annahmen ein konkretes statistisches Verfahren anwendet. Die zentrale Frage, die dann mittels des simulationsbasierten Zuganges betrachtet wird, lautet: Wie *robust* ist das angewendete statistische Verfahren gegenüber Verletzungen der Voraussetzungen bzw. Annahmen? Als Beispiel einer Studie, die mittels Simulationen die Fragen der Kovariaten- und Modellselektion untersucht, wurde bereits in Abschnitt 5.3.1 die Untersuchung von McCaffrey et al.

(2004) angeführt.

Der Vorteil dieses Ansatzes besteht darin, dass man die wahren Zusammenhänge und stochastischen Abhängigkeiten der Variablen kennt, denn man legt diese zu Beginn einer Simulationsstudie fest. Folglich hat man eine fixe Referenz – wie z. B. den zuvor festgelegten wahren kausalen Effekt –, die zur Quantifizierung des Fehlers (Bias) dient. Durch die systematische Variation der wahren Parameter (Populationskennwerte) – dem Design einer Simulationsstudie – lassen sich schließlich Aussagen über die Robustheit eines statistischen Verfahrens ableiten. Damit lassen sich im Rahmen von Simulationsstudien auch Aussagen über die Schätzung kausaler Effekte ableiten, da eine *kausale Referenz bzw. Benchmark* vorhanden ist, indem zu Beginn der Studie der wahre kausale Effekt festgelegt wird.

Dem Vorteil dieses Verfahrens steht der Nachteil gegenüber, dass die Festlegung der Populationsparameter in einer Simulationsstudie stets mit Hypothesen über die wahren Zusammenhänge in einer konkreten empirischen Anwendung einhergeht. Dies ist insbesondere im Kontext von Vergleichsarbeiten – bzw. bei Schulleistungsstudien im Allgemeinen – eine schwierige Aufgabe, da im schulischen Kontext eine Vielzahl von Variablen eine Rolle spielen. So unterscheiden sich bspw. die Bundesländer nicht nur hinsichtlich der schulischen Bedingungen (Gliederung des Schulsystems in verschiedene Schulformen, Übergangsregelungen in die Sekundarstufe etc.), sondern auch in Bezug auf die Zusammensetzung der Schülerschaft substanziell voneinander. Ein hypothetisches Beispiel soll dies verdeutlichen: In einer Simulationsstudie soll untersucht werden, welchen Einfluss die Kovariate Sozioökonomischer Status (SES) auf die lehrerspezifischen Effektschätzungen eines Value-Added Modells hat. Abhängige Variable sei die Mathematikleistung der Schüler. Dabei wird die Stärke des Zusammenhangs zwischen SES und Mathematikleistung der Schüler sowie zwischen SES und Treatment-Zuweisung systematisch variiert. Treatment-Zuweisung bedeutet hier die Zuweisung eines Schülers mit einem bestimmten SES-Wert zu einem spezifischen Lehrer. Außerdem werden die lehrerspezifischen Effekte festgelegt, die dann zum Vergleich mit den aus den simulierten Daten geschätzten Effekten verglichen werden (*Benchmarking*). Die Simulationsstudie ermöglicht Aufschluss darüber, welche Konsequenzen die Nichtberücksichtigung der angenommenen Zusammenhänge zwischen SES und den weiteren Variablen im Modell für die Parameterschätzungen haben. Da jedoch in jedem Bundesland die empirische Verteilung des Merkmals SES sehr unterschiedlich ist, lassen sich diese Aussagen nicht generalisieren. Zudem unterscheidet sich die Vertei-

lung des SES nicht nur zwischen den Bundesländern, sondern zusätzlich innerhalb eines Bundeslandes zwischen den einzelnen Schulen und Klassen (vgl. z. B. Bonsen et al., 2010). Auch könnten weitere Abhängigkeiten, die in der Simulation nicht berücksichtigt wurden (bspw. der Zusammenhang zwischen SES und Schulform), die Ergebnisse konfundieren. Welche Auswirkungen die Nichtberücksichtigung der Variable SES in einer konkreten empirischen Anwendung hat, kann nur dann beantwortet werden, wenn die wahren Zusammenhänge im Rahmen der Simulation adäquat abgebildet werden können.

Simulationsstudien sind also vor allem dann sinnvoll, wenn man belastbare Annahmen bzw. Hypothesen über plausible Verteilungs- und Zusammenhangsstrukturen der Zufallsvariablen besitzt. Diese lassen sich im Rahmen empirischer Studien gewinnen.

5.3.3 Der empirische Zugang

Im Gegensatz zum analytischen und simulationsbasierten Ansatz kennt man im Rahmen empirischer Studien die wahren Parameter (Populationskennwerte) nicht. Jedoch lassen sich diese unter bestimmten allgemeinen Annahmen über die Beobachtungen in einer konkreten Stichprobe schätzen[2]. Wichtig ist an dieser Stelle die Anmerkung, dass sich im Rahmen empirischer Studien lediglich Hypothesen bspw. über die Verteilung einer Zufallsvariablen ableiten lassen. Diese Hypothesen lassen sich jedoch nicht verifizieren, sondern allenfalls falsifizieren (Falsifikationismus; vgl. Popper, 1934/2005). Diese wissenschaftstheoretische Anmerkung mag zunächst überflüssig erscheinen, da diese spätestens seit Popper (1934/2005) zum *Common Ground* des empirisch-wissenschaftlichen Denkens und Handelns gehört. Ich betone diesen formal-logischen Gedanken an dieser Stelle vor allem deshalb, da er entscheidend ist für die Möglichkeiten und Grenzen des Erkenntnisgewinns, d. h. der Interpretation und Generalisierbarkeit von Ergebnissen der vorliegenden Arbeit sowie weiterer Arbeiten zu diesem Thema.

Kausale Benchmark

Hinsichtlich der kausalen Benchmark, die sowohl im Rahmen des analytischen als auch des simulationsbasierten Zuganges *per definitionem* gegeben ist, muss beim empirischen Zugang nach dem Design der Studie differenziert werden: Während randomi-

[2] Von zentraler Bedeutung ist hier die *i.i.d.-Annahme*, d. h. die Annahme unabhängig und identisch verteilter (*independent and identically distributed*; *i.i.d.*) Beobachtungen.

sierte Experimente die kausale Interpretation der geschätzten Effekte ermöglichen[3], ist dies in quasi-experimentellen Designs und Beobachtungsstudien – wie den im Rahmen dieser Arbeit betrachteten Vergleichsarbeiten – nur unter Gültigkeit bestimmter Annahmen möglich (vgl. Kapitel 3, Abschnitt 3.4.1). Jedoch gibt es versuchsplanerische Möglichkeiten, die kausale Interpretation im Rahmen quasi-experimentellen Designs zu gewährleisten: Mittels eines speziellen Designs, den sog. *within-study comparisons* (z. B. LaLonde, 1986; Shadish, Clark & Steiner, 2008), ist es möglich, eine kausale Benchmark zu etablieren. Zur Illustration dieses Designs dient nachfolgend die Studie von Steiner et al. (2010), deren Ergebnisse gleichfalls auf die Bedeutung der Kovariaten- und Modellselektion hinweisen. Steiner et al. (2010) reanalysierten Daten, die mittels eines *within-study designs* erhoben worden waren. Dabei wurden die Probanden zunächst zufällig einer von zwei Gruppen zugeteilt. Die Probanden der ersten Gruppe nahmen anschließend an einem randomisierten Experiment teil, d. h., sie wurden wiederum zufällig einer von zwei Treatment-Bedingungen zugewiesen. Randomisierte Experimente werden häufig als *Goldstandard* im Rahmen der Schätzung kausaler Treatment-Effekte bezeichnet (vgl. Rubin, 2008a), da eine gelungene Randomisierung die Unverfälschtheit der Prima-Facie-Effekte impliziert (vgl. Kapitel 3, Abschnitt 3.4.1). Die Probanden der zweiten Gruppe hingegen konnten sich selbst einer der beiden Treatment-Gruppen[4] zuordnen, wobei hier von den üblichen Selbstselektionseffekten im Rahmen quasi-experimenteller Designs und Beobachtungsstudien ausgegangen werden muss. Die Treatment-Effekte wurden anschließend anhand der Daten aus dem quasi-experimentellen Design mittels verschiedener Adjustierungsmodelle (Modell 1: Propensity-Score Stratifizierung; Modell 2: Propensity-Score ANCOVA; Modell 3: Propensity-Score Gewichtung; Modell 4: ANCOVA) und mittels verschiedener Kovariatensets (u. a. demographische Variablen und Prätest-Maße) geschätzt. Die Ergebnisse der verschiedenen Adjustierungsmodelle wurden dann mit dem Ergebnis des randomisierten Experiments, das als kausale Benchmark herangezogen werden kann, verglichen. Neben der zentralen Bedeutsamkeit des Vorwissens als Kovariate

[3] Einschränkend sei hier angemerkt, dass randomisierte Experimente nur dann die kausale Interpretation der Effektschätzungen ermöglichen, falls die Randomisierung gelungen ist. In diesem Fall spricht man auch von einer sog. *happy randomization*. Führt man also eine Randomisierung durch, so bedarf es anschließend stets auch der Prüfung, ob diese gelungen ist.

[4] Wichtig ist, dass es sich im Rahmen von *within-study comparisons* sowohl im randomisierten Experiment als auch in der Beobachtungsstudie jeweils um identische Treatment-Bedingungen handelt. In dieser Untersuchung handelte es sich bei den beiden Treatment-Bedingungen um ein Mathematik- und ein Vokabel-Training. Die Daten stammen aus einer früheren Studie von Shadish et al. (2008).

zeigt die Studie weiterhin, dass die Kovariatenselektion bedeutsamer ist als die Modellselektion. Einschränkend muss hier angemerkt werden, dass sich die Studie von Steiner et al. (2010) auf einen speziellen Treatment-Effekt bezieht: Es wurde der Effekt eines Mathematik-Trainings im Vergleich zu einem Vokabel-Training (und vice versa) untersucht. Zudem lag der Fokus des Interesses auf dem *ACE* und nicht auf dem *ACE on the treated*. Die Ergebnisse dieser Studie lassen sich somit nur eingeschränkt auf den im Rahmen dieser Arbeit betrachteten Schulleistungskontext und die Analyse von Unterrichtseffekten generalisieren.

Der Vorteil dieser versuchsplanerischen Technik liegt auf der Hand: Within-study designs bieten die einzigartige Möglichkeit zur Quantifizierung der Abweichung vom kausalen Effekt (d. h. des Bias) im Rahmen empirischer Beobachtungsstudien. Diesem Vorteil steht jedoch der Nachteil gegenüber, dass sich auch dieses Design – ebenso wie ein randomisiertes Experiment – nicht immer anwenden lässt: Insbesondere bei Large Scale Assessements in der Schulleistungsforschung ist ein solches Design praktisch nicht realisierbar.

Anknüpfend an die beschriebenen Einschränkungen hinsichtlich des Designs sowie der Generalisierbarkeit der Ergebnisse von Steiner et al. (2010) werden im Folgenden empirische Befunde zu den beiden Teilfragen fairerer Vergleiche im Kontext von Schulleistungsuntersuchungen dargestellt. Diese beziehen sich auf Daten aus Beobachtungsstudien, d. h. es steht keine kausale Benchmark zur Verfügung.

Wissenschaftliche Befunde

Im Kontext von Value-Added Modellen in den USA und auch in der europäischen Schuleffektivitätsforschung finden sich zahlreiche empirische Studien, die sowohl die Kovariaten- als auch Modellselektion betrachten. Im Folgenden werden einige dieser Untersuchungen – geordnet nach dem jeweiligen inhaltlichen Fokus – exemplarisch dargestellt[5].

Kovariaten- und Modellselektion. So führen beispielsweise Ballou, Sanders und Wright (2004) einen empirischen Modellvergleich durch, um den Einfluss der Kova-

[5]Dabei sei an dieser Stelle darauf hingewiesen, dass hinsichtlich der Auswahl der Studien keinesfalls der Anspruch auf Vollständigkeit besteht. Dies würde über den Rahmen der vorliegenden Arbeit hinausgehen. Im Folgenden werden beispielhaft v. a. Studien aus der amerikanischen VAM-Literatur berichtet, aber auch der Verweis auf die europäische Schuleffektivitätsliteratur wird gegeben.

riatenselektion zu analysieren. Sie verwenden in ihren Analysen das *Tennessee Value-Added Assessment System* (TVAAS; Sanders & Horn, 1994). Das TVAAS ist ein vergleichsweise sparsames Value-Added Modell, bei dem bis auf den Vortest keine weiteren Kovariaten enthalten sind. Die Autoren finden, dass die Hinzunahme von Kovariaten auf Schülerebene (z. B. SES und weitere demographische Faktoren) in das Modell keine bedeutsamen Unterschiede in den geschätzten Lehrereffekten verursacht. Als Vergleichskriterium wurden hier u. a. die Korrelationen der aus den verschiedenen Modellen resultierenden Effektschätzungen betrachtet, die – mit wenigen Ausnahmen – über alle Fächer und Klassenstufen der untersuchten Stichprobe größer als $r = .90$ waren.

Auch Tekwe et al. (2004) führen einen empirischen Modellvergleich verschiedener Value-Added Modelle durch, wobei sich die betrachteten Modelle hinsichtlich der Kovariatenselektion als auch der gewählten Parametrisierung[6] (Modellselektion) unterscheiden. Auch bei dieser Studie werden die Korrelationen der aus den verschiedenen Modellen resultierenden Effektschätzungen als Kriterium betrachtet. Im Ergebnis zeigt sich, dass die Selektion der Kovariaten einen größeren Einfluss auf die Effektschätzungen hat als die Parametrisierung.

Kovariatenselektion: Vorwissen. Es besteht ein allgemeiner Konsens, dass der Prätest bzw. das Vorwissen eine der bedeutsamsten Determinanten schulischer Leistung ist (vgl. z. B. Renkl, 1996). So konnte bspw. gezeigt werden, dass das fachspezifische Vorwissen eine bessere prädiktive Erklärungskraft als die Intelligenz bei der Vorhersage schulischer oder auch beruflicher Leistungen besitzt: In verschiedenen Studien, in denen Vorwissen und Intelligenz simultan als Prädiktoren verwendet wurden, erwies sich das fachspezifische Vorwissen als der stärkere Prädiktor zur Vorhersage von Testleistungen in der gleichen Testdomäne (Ceci & Liker, 1986; Schneider & Bjorklund, 1992; Weinert & Helmke, 1995).

Weitere empirische Evidenz, dass das fachspezifische Vorwissen eine der zentralen Kovariaten darstellt, findet sich bei Hedges und Hedberg (2007). In dieser Studie wurden Schulleistungsdaten aus den USA in den Domänen Mathematik und Lesen (Englisch) mittels verschiedener Modelle analysiert. Dabei wurden insgesamt vier Model-

[6] Mit *Parametrisierung* ist gemeint, dass eine bestimmte Gleichung aufgestellt wird, die den Zusammenhang der im Modell enthaltenen Variablen mit einer bestimmten Anzahl von zu schätzenden Parametern beschreibt. Somit ist die Wahl der Parametrisierung gleichbedeutend mit der Modellwahl.

le verglichen, die jeweils unterschiedliche Kovariatensets enthielten (Modell 1: ohne Kovariaten; Modell 2: demographische Kovariaten (u. a. Geschlecht, SES); Modell 3: fachspezifisches Vorwissen; Modell 4: fachspezifisches Vorwissen & demographische Kovariaten). Das fachspezifische Vorwissen sowie die demographischen Variablen wurden sowohl auf Individualebene als auch Schulebene (d. h. der Mittelwert der jeweiligen Variable über alle Schüler einer Schule) in das Modell aufgenommen. Sowohl für Mathematik als auch für Lesen fanden die Autoren einen signifikanten Effekt des fachspezifischen Vorwissens, dessen Ausmaß in beiden Domänen vergleichbar war. Außerdem zeigte sich folgendes Ergebnis: „In general, demographic characteristics explain little additional variance (at either the student or the school level) beyond what is explained by the pretest, and thus their inclusion in analysis models does not appear to be useful if pretest scores are available“ (Hedges & Hedberg, 2007, S. 69). Zudem wurde neben dem fachspezifischen Vorwissen auch die Bedeutung eines spezifischen Kompositionsmerkmals – das Fähigkeits- und Leistungsniveau der Schülerschaft einer Schule[7] – als relevante Kovariate deutlich. Ähnliche Befunde liefern auch die im nachfolgenden Abschnitt dargestellten Studien.

Kovariatenselektion: Kompositionsmerkmale. Leckie und Goldstein (2009) analysierten die Schätzung von Schuleffekten im Rahmen der englischen Key Stage Tests, die in Form von League Tables veröffentlicht werden und die Schulwahl von Eltern schulpflichtiger Kinder unterstützen sollen (vgl. Kapitel 4). Die Autoren führten u. a. einen Modellvergleich zwischen einem klassischen Value-Added Model (VAM) und einem Contextual Value-Added Modell (CVA) durch. Während in dem VAM das Vorwissen der Schüler sowie individuelle Schülermerkmale berücksichtigt werden, auf die eine Schule keinen Einfluss hat, wird bei dem CVA zusätzlich der Mittelwert und die Standardabweichung der Testleistungen der Schüler einer Schule im vorangegangenen Key Stage Test in das Analysemodell einbezogen. Im Ergebnis zeigte sich, dass die zusätzliche Berücksichtigung der leistungsmäßigen Komposition einer Schule zu einer substanziellen Veränderung der Schuleffekte sowie der Rangposition vieler Schulen führte. Zudem betrug die Korrelation der Schuleffekte aus den beiden Adjustierungs-

[7] Baumert et al. (2006) unterscheiden fünf zentrale Dimensionen von Kompositionsmerkmalen (vgl. auch Kapitel 4, Abschnitt 4.2): die soziokulturelle Zusammensetzung, die Konzentration sozialer Risikofaktoren durch belastende Familienverhältnisse, die ethnisch-kulturelle Zusammensetzung, die Konzentration lernbiographischer Belastungsfaktoren und das Fähigkeits- und Leistungsniveau der Schülerschaft.

modellen $r = .83$. Zwar ist dies eine starke Korrelation (Cohen, 1988); relativ zu Ergebnissen aus ähnlichen Modellvergleichen, bei denen die Korrelationen stets größer als $r = .90$ waren, ist diese jedoch vergleichsweise niedrig. Zu einem vergleichbaren Ergebnis kommen Benton et al. (2003): Auch diese Untersuchung von Schuleffektschätzungen verschiedener Modelle (VAM vs. CVA) im Kontext englischer Key Stage Tests zeigte, dass die Inklusion vom Kompositionsmerkmalen die Rangordnung der Schulen deutlich verändert. Benton et al. (2003) fanden Korrelation zwischen $r = .71$ (Key Stage 3) und $r = .94$ (Key Stage 4). Wie bereits in Kapitel 4 dargestellt, plädieren Leckie und Goldstein (2009) dafür, dass im Rahmen der Ergebnisrückmeldung in englischen League Tables keine Kompositions- bzw. Kontexteffekte berücksichtigt werden sollten, da diese Ranglisten in erster Linie der Information von Eltern für die Schulwahl dienen.

Auch die Ergebnisse von Timmermans et al. (2011) lassen sich in diese Befundlage einfügen. Neben einer Klassifikation verschiedener Typen von VAM (vgl. Kapitel 4, Abschnitt 4.2.3), bei denen die Unterschiede hinsichtlich der Interpretation der resultierenden Effektschätzungen erörtert werden, führen die Autoren gleichfalls einen empirischen Modellvergleich durch. Die zur Reanalyse verwendeten Daten stammen aus der *Cohort Study in Secondary Education* – einer niederländischen, nationalen Längsschnittstudie in der Sekundarstufe. Auch hier finden sich deutliche Kontexteffekte, die die Schuleffektschätzungen beeinflussen. Timmermans et al. (2011) zeigen, dass „... the choice of model has a large impact on the individual schools if this model would be used in an accountability system“ (S. 409). Die Korrelation der Effektschätzungen betrugen in Abhängigkeit vom Schultyp $r = .90$ bzw. $r = .94$. Die Übereinstimmung der Rangklassifikationen beim Vergleich des VAM mit dem CVA waren mit $\kappa = .72$ und $\kappa = .73$ jedoch deutlich geringer. Im Gegensatz zu Leckie und Goldstein (2009) argumentieren sie jedoch für die Inklusion von Kompositionsmerkmalen, da das Ziel dieser Modelle in der Identifikation potenziell unterstützungsbedürftiger Schulen und nicht in der Schulentwicklung sowie – wie bei Leckie und Goldstein (2009) und den englischen League Tables – in der elterlichen Schulwahl liegt.

Kovariatenselektion: Domänenspezifität. Auch Briggs und Domingue (2011) führten eine Untersuchung zur Bedeutung der Kovariatenselektion für die Effektschätzungen aus Value-Added Modellen durch. Neben der Bedeutung des Vorwissens sowie der Komposition der Schülerschaft, weisen die Befunde jedoch zusätzlich auf die *Do-*

mänenspezifität der Ergebnisse aus Value-Added Modellen hin. Domänenspezifität bedeutet in diesem Kontext, dass sich der Einfluss der Kovariaten- bzw. Modellselektion auf die Effektschätzungen zwischen verschiedenen Testdomänen wie bspw. Mathematik und Lesen unterscheidet. Da die Studie große Ähnlichkeiten zu dem in Rahmen der vorliegenden Arbeit gewählten Vorgehen aufweist, soll sie nachfolgend näher erläutert werden.

Die Autoren reanalysierten Daten aus amerikanischen Grundschulen im *Los Angeles Unified School District*, die sog. LAUSD-Daten. Dabei handelt es sich um Schulleistungsdaten in den Domänen Mathematik und Lesen (Englisch), die mittels standardisierter Tests (California Standardized Test) in den Schuljahren 2002/2003 bis 2008/2009 erhoben worden waren. Die Stärke des LAUSD-Datensatzes ist, dass dieser längsschnittliche Daten (Klassenstufe 2 bis 5) von sechs Schülerkohorten enthält. Das Ziel der Untersuchung von Briggs und Domingue (2011) bestand zum einen darin, die Ergebnisse eines Value-Added Modells (*Los Angeles Value-Added Model*; LAVAM) zur Schätzung der Schul- und Lehrereffekte anhand der LAUSD-Daten zu replizieren, die zuvor von der Los Angeles Times veröffentlicht worden waren (vgl. Felch, Song & Smith, 2010). Die Los Angeles Times klassifizierte die Lehrer in fünf Kategorien[8] entsprechend der Quintile der Verteilung der Lehrereffektschätzungen. Die Replikation der Ergebnisse der ursprünglichen LAVAM-Analyse scheiterte. Ursächlich dafür sind möglicherweise Unterschiede zwischen den jeweils verwendeten Stichproben. So ließen sich die Kriterien, die in der ursprünglichen LAVAM-Analyse zum Ausschluss von Untersuchungseinheiten aus dem Datensatz und somit der Analyse führten, aufgrund mangelhafter Dokumentation nicht nachvollziehen. Zum anderen wurde die Sensitivität der Effektschätzungen betrachtet, d. h. folgende Forschungsfrage wurde untersucht: Wie *sensitiv* bzw. unterschiedlich sind die Effektschätzungen, wenn weitere Kovariaten in das ursprüngliche LAVAM aufgenommen werden? Das *alternative Value-Added Modell* (altVAM) enthielt folgende zusätzliche Kovariaten:

- *Longer history of a student's test performance*: Zusätzlich zu den Testwerten der vorangegangenen Klassenstufe 4 wurden auch die fachspezifische Vortestwerte der Klassenstufen 2 und 3 auf Schülerebene aufgenommen.

- *Peer influence*: Weiterhin wurde der Mittelwert der Testwerte aus der vorangegan-

[8] Die fünf Kategorien, die die Effektivität der Lehrer beschreiben sollen, sind *least effective*, *less effective*, *average*, *more effective* und *most effective*.

genen Klassenstufe 4 über die Schüler der betrachteten Klasse in Klassenstufe 5 berechnet und in das altVAM als Kovariate aufgenommen.

- *School-level factors*: Es wurde zudem ein Indikator verwendet, der die Platzierung der jeweiligen Schule innerhalb des *California School Similarity Rank* angibt. Dieser ist ein Maß für die Ähnlichkeit von Schulen hinsichtlich ihrer demographischen Zusammensetzung[9].

Die Ergebnisse beider Modelle – des ursprünglichen LAVAM und des altLAVAM – wurden anschließend miteinander verglichen. Als Kriterium für den Modellvergleich wurde u. a. die Korrelation der aus beiden Modellen geschätzten fachspezifischen Lehrereffekte berechnet. Außerdem wurde die Veränderung der Lehrerklassifikation (Quintil-Ranking der Effektschätzungen; von *least effective* bis *most effective*), die sich unter Verwendung des altVAM anstelle des LAVAM ergab, betrachtet. Auch dies erfolgte domänenspezifisch, d. h. jeweils separat für die Bereiche Mathematik und Lesen. Dabei findet sich hinsichtlich sämtlicher Kriterien des Modellvergleichs eine bedeutsame Variabilität bzw. Sensitivität der Effektschätzungen hinsichtlich der Kovariatenselektion – sowohl für Mathematik als auch für Lesen. Die Ergebnisse bestätigen die Bedeutsamkeit der Kovariaten Vorwissen sowie von Kompositionsmerkmalen von Schulen und Klassen. Briggs und Domingue (2011) diskutieren an dieser Stelle jedoch auch kritisch das Fehlen einer kausalen Benchmark im Rahmen ihrer Analyse und somit die gleichfalls eingeschränkte Validität der Effektschätzungen aus dem altVAM hinsichtlich ihrer kausalen Interpretierbarkeit. Zudem führen die Autoren Unterschiede hinsichtlich der inhaltlichen Interpretation der Effektschätzungen an, die durch die Aufnahme zusätzlicher Kovariaten in das Modell resultieren (vgl. auch Kapitel 4, Abschnitt 4.2.1). Des Weiteren ist die Sensitivität der Effektschätzungen domänenspezifisch: Das Ausmaß des Einflusses der verschiedenen Modelle – LAVAM vs. altVAM – auf die lehrerspezifischen Effektschätzungen ist beträchtlich stärker für den Bereich Lesen als hinsichtlich der Mathematik-Testwerte. Während für die Mathematik-Outcomes 60.8% der Lehrer die gleiche Effektivitätsklassifikation bei dem Modellvergleich erhielten, behielten lediglich 46.4% der Lehrer das gleiche Effektivitätsrating in der Domäne Lesen. Auch die Korrelation der lehrerspezifischen Effektschätzungen ist in Mathematik mit $r_{math} = .92$

[9] Das Vorgehen zur Berechnung dieses Indikators ist vergleichbar mit Strategie IIIc zur Bildung der Kontextgruppen, die die soziale Belastung einzelner Klassen bzw. Schulen beschreiben (vgl. Kapitel 4, Abschnitt 4.1.2).

stärker als in Lesen ($r_{reading}$ = .76). Diese Domänenspezifität hinsichtlich der Sensitivität der Lehrereffektschätzungen ist möglicherweise darauf zurückzuführen, dass der Bereich Lesen im Vergleich zu Mathematik eine schulunabhängigere Domäne ist (vgl. Baumert, Becker, Neumann & Nikolova, 2009).

5.4 Fragestellungen und Hypothesen

Im folgenden Abschnitt werden die zentralen Fragestellungen dieser Arbeit sowie die Wahl des methodischen Zuganges dargestellt. Anschließend werden die Hypothesen spezifiziert, die im empirischen Teil dieser Arbeit einer Überprüfung an Beobachtungsdaten, d. h. an Testleistungsdaten aus Vergleichsarbeiten, unterzogen werden.

5.4.1 Offene Fragen bei Vergleichsarbeiten

Ausgehend von der Differenzierung der Problematik fairerer Vergleiche in die zwei Facetten Kovariatenselektion und Modellselektion (vgl. Abschnitt 5.2) sowie den bisherigen Forschungsbefunden (vgl. Abschnitt 5.3), geht es im Rahmen der vorliegenden Arbeit um folgende Fragestellungen:

(1) Kovariatenselektion: Welchen Einfluss hat die zusätzliche Berücksichtigung des fachspezifischen Vorwissens sowie von Klassenkompositionsmerkmalen (insbesondere des klassenspezifischen Leistungsniveaus) auf die klassenspezifischen Effektschätzungen?

(2) Modellselektion: Welchen Einfluss haben Variationen der Modellspezifikation auf die klassenspezifischen Effektschätzungen?

(3) Fachspezifität: Gibt es Unterschiede hinsichtlich des Einflusses der Kovariaten- und Modellselektion auf die klassenspezifischen Effektschätzungen zwischen verschiedenen Unterrichtsfächern, in denen Vergleichsarbeiten durchgeführt werden?

5.4.2 Methodischer Zugang: Empirische Reanalyse von Daten aus Vergleichsarbeiten

Zur Beantwortung dieser Fragen wähle ich im Rahmen der vorliegenden Arbeit den empirischen Zugang, um die konkreten Gegebenheiten (d. h. die Verteilung und Zusammenhänge der Variablen) bei Daten aus Vergleichsarbeiten bestmöglich abzubilden. Im Rahmen einer empirischen Reanalyse von Daten aus dem Projekt *Kompetenztest.de* werden verschiedene Adjustierungsmodelle auf dieselbe Datenbasis angewendet. Der Vergleich der jeweils resultierenden Ergebnisse ermöglicht Rückschlüsse hinsichtlich der Bedeutung der Variablenselektion sowie der Wahl des statistischen Modells. Diese empirische Reanalyse ist – ebenso wie die Studie von Briggs und Domingue (2011) – als eine Sensitivitätsanalyse aufzufassen: Die Sensitivität der Effektschätzungen gegenüber der Modellspezifikation und der Auswahl der Kovariaten wird analysiert, wobei insbesondere die Bedeutung der Kovariaten Vorwissen (Individualmerkmal) und Leistungsniveau der Schüler einer Klasse (Klassenkompositionsmerkmal) betrachtet wird. Da sich die Reanalyse auf Thüringer Kompetenztestdaten beziehen, wird das Adjustierungsverfahren des Projektes *Kompetenztest.de* Ausgangspunkt der empirischen Analyse sein und gleichfalls als Referenz des Modellvergleichs herangezogen.

Die nachfolgende Analyse liefert somit empirische Evidenz über die Bedeutung der Kovariatenselektion und der Modellannahmen im Kontext von Vergleichsarbeiten. Diese kann im Rahmen von Best-Practice-Entscheidungen zur Entwicklung neuer oder Weiterentwicklung bestehender Adjustierungsverfahren – insbesondere im Rahmen der Ergebnisrückmeldung von Testleistungen aus Vergleichsarbeiten – genutzt werden. Ein weiterer zentraler Nutzen der im Rahmen der vorliegenden Arbeit gewählten Herangehensweise besteht in der Wahl der Methode sowie der Beurteilungskriterien selbst. So können die Ergebnisse der empirischen Reanalyse von Daten aus dem Bundesland Thüringen selbstverständlich nicht ohne Weiteres auf Adjustierungsverfahren in anderen Bundesländern übertragen werden (vgl. Kapitel 8). Das im Rahmen dieser Arbeit gewählte Vorgehen sowie die verwendeten Beurteilungskriterien können jedoch als Richtlinien für zukünftige Analysen in diesem Anwendungskontext genutzt werden.

5.4.3 Hypothesen

Bereits in Kapitel 4 (vgl. Abschnitt 4.1.2) wurde auf eine Besonderheit der Thüringer Kompetenztestdaten verwiesen, die diese insbesondere für die Untersuchung der ersten Fragestellung (Kovariatenselektion) prädestiniert: Bisher liegen im Kontext von Vergleichsarbeiten in allen Bundesländern lediglich querschnittliche Datensätze vor und es stehen zumeist auch keine sonstigen mit dem Vorwissen assoziierten Maße – z. B. kognitive Grundfähigkeiten, die mit dem KFT im Rahmen von PISA-E 2000 erhoben wurden – der Schüler zur Verfügung. Lediglich in Thüringen ist seit dem Schuljahr 2009/2010 eine längsschnittliche Verknüpfung der Kompetenztestdaten aus den verschiedenen Erhebungsjahrgängen möglich. Die Kompetenztestergebnisse früherer Jahrgänge dienen als Maß des fachspezifischen Vorwissens. Das Leistungsniveau der Schüler einer Klasse wird nachfolgend über den Mittelwert des fachspezifischen Vorwissens aller Schüler einer Klasse operationalisiert. In Anlehnung an die empirischen Befunde bisheriger Studien (vgl. Abschnitt 5.3) wird angenommen, dass beide Kovariaten – das individuelle Vorwissen eines Schülers und das klassenspezifische Leistungsniveau – einen bedeutsamen Einfluss auf die klassenspezifischen Effektschätzungen haben.

In Abschnitt 5.3.1 wurde dargestellt, dass das in PISA-E 2000 verwendete Adjustierungsverfahren ein Spezialfall des Adjustierungsverfahrens im Projekt *Kompetenztest.de* ist (vgl. Seite 110 sowie Fiege, 2007): Während bei dem Adjustierungsmodell im Projekt *Kompetenztest.de* keine Annahmen über die funktionale Form gemacht werden, handelt es sich bei dem im Rahmen von PISA-E 2000 verwendeten linearen Modell um eine restriktivere, sparsamere Parametrisierung. Über die Plausibilität der Linearitätsannahme lässt sich – wie bereits erwähnt – im Rahmen des analytischen Zuganges keine Aussage treffen, sondern hierzu bedarf es eines empirischen Modellvergleichs. Es wird angenommen, dass die klassenspezifischen Effektschätzungen sensitiv gegenüber Veränderungen der Parametrisierung (Modellselektion) sind, wobei eine lineare Parametrisierung ohne Berücksichtigung potenzieller Interaktionen nicht ausreichend ist, den wahren Zusammenhang abzubilden[10]. Zudem wird aufgrund bisheriger Befunde (vgl. Abschnitt 5.3) angenommen, dass die Selektion der Kovariaten einen größeren Einfluss auf die Effektschätzungen hat als die Modellselektion bzw. Parametrisierung.

Vergleichsarbeiten wie die Thüringer Kompetenztests werden in den Fächern Mathe-

[10] Eine detaillierte Beschreibung der im Rahmen des empirischen Modellvergleichs verwendeten Modelle erfolgt in Kapitel 6.

matik und Deutsch durchgeführt[11]. Aufgrund der in Abschnitt 5.3 berichteten Befunde wird angenommen, dass die postulierten Einflüsse der Kovariatenselektion (individuelles Vorwissen sowie klassenspezifisches Leistungsniveau) und Modellselektion sowohl für Mathematik als auch Deutsch zu finden sind. Zwar beziehen sich die in Abschnitt 5.3 berichteten Befunde auf Domänen und nicht auf Unterrichtsfächer, jedoch umfassen die fachspezifischen Vergleichsarbeiten verschiedene Domänen. So werden im Fach Mathematik und Deutsch jeweils mindestens zwei Domänen getestet. Zudem wird angenommen, dass für beide Fächer die Selektion der Kovariaten einen größeren Einfluss auf die Effektschätzungen hat als die Parametrisierung. Aus den bisherigen Befunden lassen sich jedoch keine plausiblen Hypothesen über eine fächerspezifische Sensitivität der Effektschätzungen ableiten. Zwar zeigt sich bei Briggs und Domingue (2011), dass der Einfluss der verwendeten Modelle auf die lehrerspezifischen Effektschätzungen beträchtlich stärker für den Bereich Lesen als für Mathematik ist. Hier ist jedoch zu berücksichtigen ist, dass es sich bei den LAUSD-Daten um Schulleistungsdaten der Grundschule handelt, so dass sich diese Befunde nicht ohne Weiteres auf Daten der Sekundarstufe generalisieren lassen. Zudem können weitere Faktoren Einfluss darauf haben, ob und in welche Richtung derartige fachspezifischen Unterschiede bezüglich der Sensitivität der Effektschätzungen vorliegen. Dazu zählen u. a. die psychometrische Qualität der zugrundeliegenden Tests, die Schulabhängigkeit der betrachteten Domänen oder Unterschiede hinsichtlich Ausgangsniveau und Variabilität der Testwerte sowie dem Ausmaß der Leistungsentwicklung – um nur einige Beispiele zu nennen. Im Hinblick auf die Fachspezifität kann daher nicht von der Generalisierbarkeit der Ergebnisse von Briggs und Domingue (2011) auf den Kontext von Vergleichsarbeiten ausgegangen werden. Daher postuliere ich die ungerichtete Annahme, dass sich die Sensitivität der klassenspezifischen Effektschätzungen sowohl für Mathematik als auch Deutsch zeigt bzw. über beide Fächer generalisieren lässt.

[11] In Klassenstufe 8 sowie in einigen Bundesländern auch in Klassenstufe 6 wird zudem die erste Fremdsprache (Englisch oder Französisch) getestet. Bei den Kompetenztests in Klassenstufe 8 ist die Teilnahme an dem Fremdsprachentest obligatorisch, jedoch kann frei gewählt werden, ob an dem Kompetenztest Englisch oder Französisch teilgenommen wird. Somit findet hier keine Vollerhebung statt. Aus diesem Grund werden die Fächer Englisch und Französisch im Rahmen der vorliegenden Arbeit nicht betrachtet werden.

Die dargelegten Annahmen können mit den folgenden Hypothesen geprüft werden:

Hypothese 1: Kovariatenselektion

1.1: Kovariate Vorwissen
Die zusätzliche Berücksichtigung des fachspezifischen Vorwissens – additional zu den weiteren Kovariaten – im Adjustierungsmodell des Projektes *Kompetenztest.de* führt zu einer Veränderung der klassenspezifischen Effektschätzungen.

1.2: Kovariate Klassenkomposition
Die zusätzliche Berücksichtigung der leistungsmäßigen Klassenkomposition (d. h. das durchschnittliche Leistungsniveau der Schüler einer Klasse) – additional zu den weiteren Kovariaten und zum Vorwissen – im Adjustierungsmodell des Projektes *Kompetenztest.de* führt zu einer Veränderung der klassenspezifischen Effektschätzungen.

Hypothese 2: Modellselektion
Eine sparsamere Parametrisierung führt zu einer Veränderung der klassenspezifischen Effektschätzungen: Die Annahme eines bedingt linearen Zusammenhangs (inklusive der Modellierung von Interaktionstermen) im Adjustierungsmodell des Projektes *Kompetenztest.de* ist einer sparsameren linearen Parametrisierung (ohne Interaktionsterme) vorzuziehen.

Hypothese 3: Kovariatenselektion vs. Modellselektion
Je mehr relevante Variablen im Adjustierungsmodell des Projektes *Kompetenztest.de* enthalten sind, desto geringer ist der Einfluss der Modellspezifikation auf die klassenspezifischen Effektschätzungen.

Hypothese 4: Generalisierung über Fächer
Bezüglich der Hypothesen 1 bis 3 besteht keine Fachspezifität hinsichtlich des Adjustierungsverfahrens des Projektes *Kompetenztest.de*: Die in Hypothese 1 bis 3 postulierten Zusammenhänge, d. h. die Sensitivität der Effektschätzungen gegenüber der Kovariaten- und Modellselektion, lassen sich sowohl im Fach Mathematik als auch im Fach Deutsch finden.

Design trumps analysis.
DONALD B. RUBIN (2008)

6 Methode: Empirischer Vergleich verschiedener Adjustierungsmodelle

Zum Zweck der Prüfung der im vorangegangenen Kapitel explizierten Hypothesen wurde im Rahmen einer empirischen Reanalyse von Schulleistungsdaten aus dem Projekt *Kompetenztest.de* ein Modellvergleich verschiedener Adjustierungsmodelle durchgeführt. Im Fokus der Analysen steht die Sensitivität der Effektschätzungen gegenüber der Modellspezifikation und der Auswahl der Kovariaten. Die Daten entstammen den Vergleichsarbeiten des Freistaates Thüringen, die als Kompetenztests bezeichnet werden. Im nachfolgenden Kapitel werden – nach einer Einführung in das Vorgehen bei der jährlichen Erhebung der Thüringer Kompetenztests – die verwendeten Erhebungsinstrumente und Variablen vorgestellt. Im Zentrum des folgenden Kapitels steht das methodische Vorgehen bzw. das Design des Modellvergleichs.

6.1 Die Thüringer Kompetenztests und das Projekt *Kompetenztest.de*

Die Thüringer Kompetenztests werden jährlich in der Klassenstufe 3 in den Fächern Mathematik und Deutsch sowie in den Klassenstufen 6 und 8 in den Fächern Mathematik, Deutsch und Englisch[1] durchgeführt. Für die Durchführung der Tests sind die teilnehmenden Thüringer Schulen zuständig, d. h. sowohl die Testdurchführung als auch die Testkorrektur obliegt den Lehrkräften des entsprechenden Unterrichtsfaches. Die Teilnahme ist in der dritten und achten Klasse für alle Tests obligatorisch. Dahingegen

[1] In Klassenstufe 8 kann anstatt des Kompetenztests im Fach Englisch auch der Test im Fach Französisch gewählt werden.

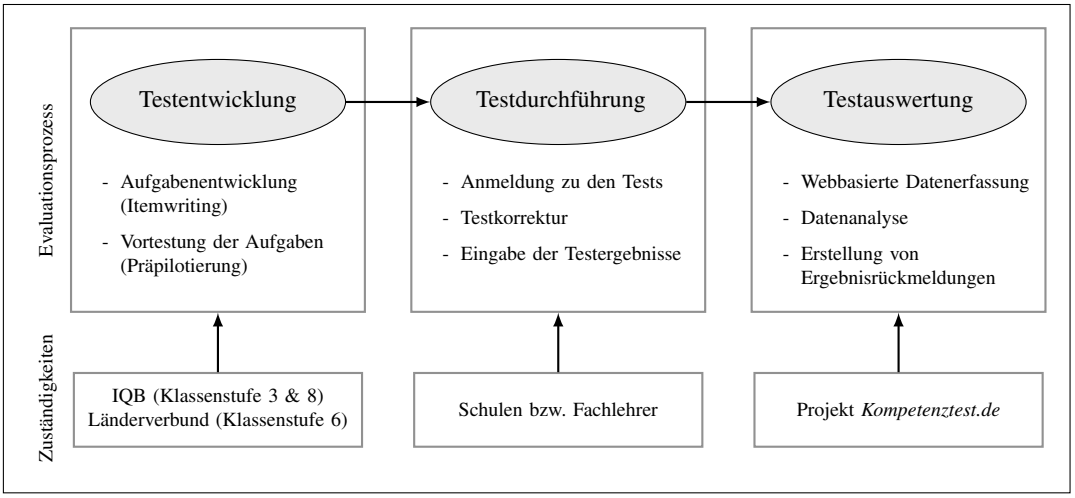

Abbildung 6.1: Zuständige Institutionen im Rahmen der Testentwicklung, Durchführung und Auswertung der Thüringer Kompetenztests

ist die Teilnahme der sechsten Klassen seit dem Schuljahr 2008/2009 wahlobligatorisch[2]. Das bedeutet, dass die Klassen lediglich an mindestens einem der Tests teilnehmen müssen (Nachtigall, 2010). Die Schulen entscheiden hier eigenverantwortlich über die Auswahl des Testfaches. Abbildung 6.1 zeigt den Ablauf des Evaluationsprozesses und die jeweils zuständigen Institutionen.

Für die Entwicklung der Thüringer Kompetenztests sind verschiedene Institutionen zuständig: Die Tests für die Klassenstufe 3 wurden bis zum Schuljahr 2008/2009 von der Projektgruppe VERA an der Universität Koblenz-Landau entwickelt. Im Schuljahr 2009/2010 wurde die Entwicklung der Tests in Klassenstufe 3 für alle Bundesländer durch das Institut für Qualitätsentwicklung im Bildungswesen (IQB) in Berlin übernommen. Das IQB ist seit dem Schuljahr 2008/2009 gleichfalls für die länderübergreifende Erstellung der Tests in Klassenstufe 8 verantwortlich. Die Entwicklung der Tests für Klassenstufe 6 hingegen erfolgt seit dem Schuljahr 2007/2008 in einem Verbund der Bundesländer Hamburg, Hessen, Mecklenburg-Vorpommern, Sachsen, Schleswig-Holstein und Thüringen (vgl. Nachtigall, 2008).

Die internetbasierte Datenerfassung, die Auswertung der Daten sowie die Erstellung der Ergebnisrückmeldungen erfolgt durch das Projekt *Kompetenztest.de* der Friedrich-

[2]Bis zum Schuljahr 2007/2008 war die Teilnahme der sechsten Klassen für alle drei Fächer – Mathematik, Deutsch und Englisch – verpflichtend.

Schiller-Universität Jena[3]. Im Rahmen der internetbasierten Anmeldung der Schüler zu den Kompetenztests durch die Lehrer der teilnehmenden Schulen wird für jeden Schüler ein Schülercode erstellt. Dadurch ist eine schülerspezifische Zuordnung der Ergebnisse seitens des Lehrers möglich, wobei gleichzeitig die Anonymität der Schüler gewährleistet wird. Eine längsschnittliche Verknüpfung der Leistungsdaten einzelner Schüler über verschiedene Erhebungsjahre seitens des Fachlehrers oder des Projektes *Kompetenztest.de* ist hingegen aus datenschutzrechtlichen Gründen ausgeschlossen. Jedoch nehmen die Schulen seit 2005 – zusätzlich zur jährlichen Anmeldung zu den Thüringer Kompetenztests – auch an dem sog. *Thüringer Schülerlängsschnitt* teil. Hierfür wurde ein eigenständiges Erhebungsprogramm entwickelt, mittels dem Schülerstammdaten erfasst und an das Thüringer Landesrechenzentrum übermittelt werden. Die Stammdaten ermöglichen die eindeutige Identifikation der Schüler und somit die längschnittliche Verknüpfung der Daten, jedoch findet keine gemeinsame Speicherung von Stammdaten und Leistungsdaten statt. Der Thüringer Schülerlängsschnitt genügt somit datenschutzrechtlichen Anforderungen, da „ ... weder die Universität Jena (Datenauswertung) noch das Thüringer Landesrechenzentrum (Treuhänder der Stammdaten aller Schüler) Leistungsdaten zusammenführen können, ohne dass dieses Vorgehen durch das Thüringer Kultusministerium autorisiert und eine Verknüpfung durch das Thüringer Landesrechenzentrum hergestellt wird“ (Nachtigall, 2010, S. 8). Mittels des Thüringer Schülerlängschnitts war im Schuljahr 2009/2010 erstmals eine längsschnittliche Verknüpfung der Leistungsdaten aus den Kompetenztests über drei Erhebungswellen möglich. Das betrifft die Kohorte Thüringer Schüler, die im Schuljahr 2004/2005 die Klassenstufe 3 besucht haben. Tabelle 6.1 zeigt die Erhebungszeitpunkte dieser Kohorte. Die Schüler dieser Kohorte haben somit im Schuljahr 2004/2005 die Kompetenztests Mathematik und Deutsch der Klassenstufe 3, im Schuljahr 2007/2008 die Kompetenztests Mathematik, Deutsch und Englisch der Klassenstufe 6 sowie schließlich im Schuljahr 2009/2010 die Kompetenztests Mathematik, Deutsch und Englisch der Klassenstufe 8 bearbeitet. Die im Folgenden dargestellten Analysen beziehen sich auf die Kompetenztests in den Fächern Mathematik und Deutsch, von denen zu allen drei Erhebungszeitpunkten Leistungsdaten vorliegen.

[3]Das Projekt *Kompetenztest.de* wurde im Jahr 2002 im Auftrag des Thüringer Ministeriums für Bildung, Wissenschaft und Kultur (TMBWK) am Lehrstuhl für Methodenlehre und Evaluationsforschung (Prof. Rolf Steyer) der Friedrich-Schiller-Universität Jena gegründet. Leiter des Projektes *Kompetenztest.de* ist Dr. Christof Nachtigall.

Tabelle 6.1: Erhebungszeitpunkte der Kohorte 2004/2005 mit drei Erhebungswellen im Thüringer Schülerlängsschnitt

Schuljahr	Klassenstufe					
	3	4	5	6	7	8
2004/2005	×					
2005/2006		—				
2006/2007			—			
2007/2008				×		
2008/2009					—	
2009/2010						×

Anmerkungen. Die Tabelle enthält ausschließlich die Erhebungszeitpunkte der Kohorte Thüringer Schüler, die im Schuljahr 2004/2005 in Klassenstufe 3 waren. Erhebungszeitpunkte sind mit × und Schuljahre ohne Erhebung sind mit — gekennzeichnet.

6.2 Erhebungsinstrumente und Variablen

Im folgenden Abschnitt werde ich die verwendeten Erhebungsinstrumente sowie die für die Analysen zentralen Variablen vorstellen. Die mittels der Kompetenztests in den Fachbereichen Mathematik und Deutsch erfassten Testleistungen werden in den nachfolgenden Analysen als abhängige Variablen dienen. Des Weiteren sind die im Rahmen der jährlichen Erhebung erfassten fachspezifischen sowie fachunspezifischen Kovariaten für den nachfolgenden empirischen Modellvergleich von elementarer Bedeutung.

6.2.1 Kompetenztests in den Fachbereichen Mathematik und Deutsch

Die Kompetenztests Mathematik und Deutsch sind standardisierte Erhebungsinstrumente, die der Erfassung des Lern- und Leistungsstandes Thüringer Schüler im Fach Mathematik und Deutsch dienen.

Kompetenztest Mathematik. Die Testaufgaben zur Erfassung der Mathematikleistung werden auf Basis der Bildungsstandards entwickelt, welche einem dreidimensionalen Kompetenzmodell zugrunde liegen (KMK, 2004b; Köller, 2010; Institut zur Qua-

Tabelle 6.2: Die drei Dimensionen des Kompetenzmodells im Fach Mathematik

Allgemeine mathematische Kompetenzen	Leitideen	Anforderungsbereiche
(K1) Argumentieren	(L1) Zahl	(I) Reproduzieren
(K2) Problemlösen	(L2) Messen	(II) Zusammenhänge herstellen
(K3) Modellieren	(L3) Raum und Form	
(K4) Darstellungen verwenden	(L4) Daten und Zufall	(III) Verallgemeinern und Reflektieren
(K5) Technisch arbeiten	(L5) Funktionaler Zusammenhang	
(K6) Kommunizieren		

Anmerkungen. Eine detaillierte Beschreibung der allgemeinen mathematischen Kompetenzen, Leitideen und Anforderungsbereiche findet sich in KMK (2004b).

litätsentwicklung im Bildungswesen, 2008). Die erste Dimension beschreibt prozessbezogene bzw. *allgemeine mathematische Kompetenzen*. Diese umfassen sechs verschiedene kognitive Operationen, die Schüler in sämtlichen Inhaltsbereichen der Mathematik anwenden. Die zweite Dimension hingegen umfasst inhaltliche mathematische Kompetenzen – die sog. *Leitideen*. Hier werden fünf verschiedene Leitideen unterschieden. Die dritte Dimension – *Anforderungsbereiche* – definiert schließlich drei verschiedene Anforderungsniveaus, welche die Komplexität der zum Lösen der Mathematikaufgaben erforderlichen Kompetenzen beschreibt. Tabelle 6.2 fasst die verschiedenen allgemeinen Kompetenzbereiche, Leitideen und Anforderungsniveaus für das Fach Mathematik zusammen, die mittels der Testaufgaben in Vergleichsarbeiten operationalisiert werden. Die Beispielaufgabe *Eisberg* soll dies illustrieren:

> Ein Eisberg verliert pro Jahr 10% seines Volumens. [...] Der Eisberg hat ursprünglich ein Volumen von 800 km^3. Wie viel Liter verliert er in einem Jahr? (Blum, Drüke-Noe, Hartung & Köller, 2006, S. 215)

Diese Aufgabe ist dem Kompetenzbereich (K5) zuzuordnen, d. h. sie erfordert die Kompetenz, mit symbolischen, formalen und technischen Elementen der Mathematik umzugehen (*Technisch arbeiten* in Tabelle 6.2). Weiterhin erfasst die Eisberg-Aufgabe die erste Leitidee *Zahl* (L1). Diese erfasst sämtliche Aspekte, „ [...] die mit Quantifizierungen zu tun haben, das heißt mit der Verwendung von Zahlen zur Beschreibung und Organisation von Situationen“ (Institut zur Qualitätsentwicklung im Bildungswesen,

2008, S. 12). Schließlich ist diese Aufgabe dem Anforderungsbereich I (*Reproduzieren*) zugeordnet, welcher sich über das Verwenden elementarer Lösungsverfahren, das direkte Anwenden von Formeln oder Symbolen bzw. das direkte Nutzen einfacher mathematischer Werkzeuge (z. B. einer Formelsammlung) auszeichnet (Institut zur Qualitätsentwicklung im Bildungswesen, 2008).

Der Kompetenztest zur Erhebung der Mathematikleistung in Klassenstufe 8 (MK8) im Schuljahr 2009/2010 umfasste Aufgaben aus allen sechs allgemeinen mathematischen Kompetenzen, allen fünf Leitideen, und allen drei Anforderungsbreichen. Der Test wurde am 04. März 2010 durchgeführt und bestand aus insgesamt 35 Teilaufgaben bzw. Items, die von den Schülern entweder gelöst oder nicht gelöst werden konnten (vgl. Tabelle 6.3). Die Korrektur der Testaufgaben erfolgte durch die jeweiligen Fachlehrer, wobei jede richtig gelöste Aufgabe mit einem Punkt bewertet wurde. Nicht oder falsch gelöste Aufgaben wurden mit 0 Punkten bewertet. Der Testwert eines Schülers ist dann der Summenscore über alle Items, wobei hier maximal 35 Punkte (und minimal 0 Punkte) erreicht werden konnten. Im Anschluss an die Korrektur erfolgte die Eingabe der Testergebnisse – wiederum seitens der Fachlehrer – pro Schüler und Item über das Schulportal auf der Webseite des Projektes *Kompetenztest.de* (vgl. Nachtigall, 2010).

Kompetenztest Deutsch. Die Testaufgaben zur Erfassung der Deutschleistung werden ebenfalls in Anlehnung an die Bildungsstandards entwickelt. Im Fachbereich Deutsch werden in den Bildungsstandards für die Sekundarstufe I vier Kompetenzbereiche unterschieden: (a) Sprache und Sprachgebrauch untersuchen, (b) Sprechen und Zuhören, (c) Schreiben und (d) Lesen, d. h. mit Texten und Medien umgehen (KMK, 2004a). Abbildung 6.2 fasst diese vier Bereiche in einem schematischen Basismodell zusammen. Der erste Kompetenzbereich (*Sprache und Sprachgebrauch untersuchen*) steht dabei in Beziehung zu jedem der drei weiteren Bereiche, da es sich hierbei um die Voraussetzung für gelingende kommunikative Kompetenzen handelt. Für jeden der drei weiteren Bereiche wurde auf Basis der Bildungsstandards ein Kompetenzmodell entwickelt, die durch je fünf Kompetenzstufen beschrieben werden: Die einzelnen Kompetenzstufen werden mittels didaktisch nachvollziehbarer Beschreibungen der Kompetenzen charakterisiert, die mit der jeweiligen Stufe einhergehen.

Sprechen und Zuhören werden in den Bildungsstandards zu einem gemeinsamen Kompetenzbereich zusammengefasst. Dieser Bereich wird in den Bildungsstandards wiederum in verschiedene Subkomponenten differenziert. Im Kontext von Schulleis-

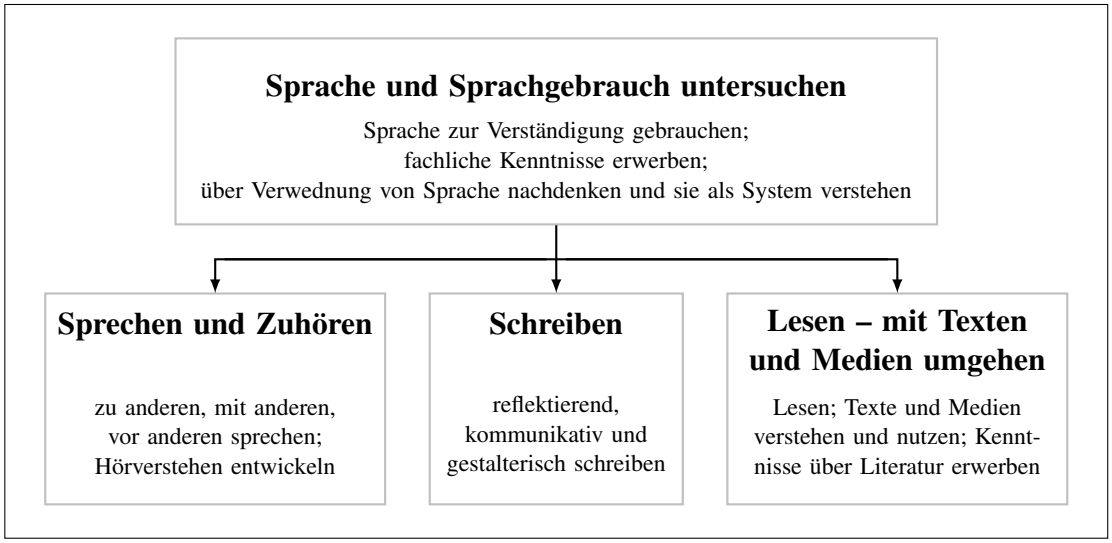

Abbildung 6.2: Kompetenzbereiche in den Bildungsstandards für den Fachbereich Deutsch in der Sekundarstufe I (in Anlehnung an KMK, 2004b)

tungsuntersuchungen sind jedoch aus testökonomischen Gründen nicht alle Teilkompetenzen operationalisierbar. So sind Aufgaben zu Subkomponenten wie bspw. zu anderen sprechen, vor anderen sprechen oder szenisches Spielen nicht ohne Weiteres im Klassenverband durchführbar. Bei der Messung der Bildungsstandards werden daher lediglich Aufgaben zum verstehenden Zuhören – als eine der Subkomponenten des Kompetenzbereichs Sprechen und Zuhören – entwickelt[4]. Beim dritten Kompetenzbereich *Schreiben* steht die Orthografie, d. h. richtiges Schreiben, im Fokus der Aufgabenentwicklung. Zur Erfassung der Rechtschreibkompetenz wurden unterschiedliche Instrumente entwickelt, wobei insbesondere Lückentextdiktate einen zentralen Aufgabentyp darstellen. Für den Kompetenzbereich *Lesen* wird in den Bildungsstandards eine große Bandbreite von Teilkompetenzen formuliert, die jedoch – ebenso wie bei den anderen Kompetenzbereichen – nicht alle im Rahmen von Schulleistungsuntersuchungen operationalisierbar sind. Im Rahmen der Messung der Bildungsstandards werden zu

[4]Die fünf Kompetenzstufen bzw. Niveaus des Kompetenzstufenmodells der Zuhörkompetenz lauten: (a) Wiedererkennen und Erinnern prominenter Einzelinformation (Niveau I), (b) benachbarte Informationen miteinander verknüpfen und den Text genrespezifisch zuordnen (Niveau II), (c) verstreute Informationen miteinander verknüpfen, der Vorlage paraverbale Informationen abgewinnen und den Text ansatzweise im Ganzen erfassen (Niveau III), (d) auf der Ebene des Textes wesentliche Zusammenhänge erkennen, die Gestaltung reflektieren und versteckte Einzelinformationen erinnern (Niveau IV) und schließlich (e) Interpretieren, Begründen, Bewerten und anspruchsvolle Erinnerungsleistungen (Niveau V).

folgenden Subkomponenten des Leseverstehens Aufgaben entwickelt: (a) Lesestrategien kennen und anwenden, (b) literarische Texte verstehen und nutzen sowie (c) Sach- und Gebrauchstexte verstehen und nutzen. Dabei liegt bei der Aufgabenentwicklung der Schwerpunkt auf den beiden zuletzt genannten Teilkompetenzen.

Der Kompetenztest im Schuljahr 2009/2010 zur Erhebung der Deutschleistung in Klassenstufe 8 (DK8) umfasste Testaufgaben aus den Bereichen Zuhören und Leseverstehen. Die Testdurchführung erfolgte am 24. Februar 2010. Der Testwert eines Schülers ist wiederum der Summenscore über alle Items, wobei hier maximal 69 Punkte (und minimal 0 Punkte) erreicht werden konnten (vgl. Tabelle 6.3). Im Anschluss an die Korrektur gaben die Fachlehrer die Testergebnisse pro Schüler und Item über das Schulportal auf der Webseite des Projektes *Kompetenztest.de* ein (vgl. Nachtigall, 2010).

6.2.2 Kovariaten

Neben den Leistungsdaten wurden weiterhin verschiedene Schülermerkmale erfasst, die bei der Berechnung des korrigierten Landesmittelwertes berücksichtigt wurden (vgl. Nachtigall, 2010, S. 34). Diese Variablen, die *Kovariaten*, sind gemeinsam mit der Mathematik- und Deutschleistung in Klassenstufe 8 in Tabelle 6.3 aufgelistet. Dabei lassen sich *fachspezifische* Kovariaten, deren Werte sich zwischen den Fächern unterscheiden können, von *fachunspezifischen* Kovariaten differenzieren. Letztere werden unabhängig von dem Unterrichtsfach erhoben und sowohl im Rahmen der Adjustierung im Fach Mathematik als auch im Fach Deutsch verwendet.

Fachspezifische Kovariaten

Die folgenden Variablen werden separat sowohl im Fach Mathematik als auch im Fach Deutsch erhoben und zur Berechnung des jeweils fachspezifischen korrigierten Landesmittelwertes verwendet[5]: die Diagnose besonderer Lernschwierigkeiten bzw. sonderpädagogischer Förderbedarf (BLSF: 0 = *kein Förderbedarf*, 1 = *Förderbedarf*) und die Anzahl der Bücher im Elternhaus. Diese sog. *Bücherfrage* hat sich als geeigneter Indi-

[5]Die Variablen tragen jeweils den gleichen Namen, sind jedoch zusätzlich mit fachspezifischen Endungen gekennzeichnet. Alle Variablen, die spezifisch für das Fach Mathematik sind, wurden mit der Endung *M* versehen. Dementsprechend sind Variablen, die spezifisch für das Fach Deutsch sind, mit der Endung *D* gekennzeichnet.

Tabelle 6.3: Übersicht der Variablen

Variable	Beschreibung	Ausprägungen und Wertelabel
Variablen im Fach Mathematik		
MK8	Mathematikleistung Klasse 8	0 – 35 (Summenscore)
MK6	Mathematikleistung Klasse 6	0 – 28 (Summenscore)
MK3	Mathematikleistung Klasse 6	0 – 24 (Summenscore)
BLSF.M	Förderbedarf in Mathematik	0 = *kein Förderbedarf*, 1 = *Förderbedarf*
SES.M	SES[a] der Mathematik-Klasse	ordinal[b]; Kategorien 1 – 4
SART.M	Schulart der Mathematik-Klasse	1 = *Förderschule*, 2 = *Hauptschule*, 3 = *nicht-differenziert*[c], 4 = *Realschule*, 5 = *Gymnasium*
Variablen im Fach Deutsch		
DK8	Deutschleistung Klasse 8	0 – 69 (Summenscore)
DK6	Deutschleistung Klasse 6	0 – 96 (Summenscore)
DK3L	Deutschleistung Klasse 3 (Lesen)	0 – 15 (Summenscore)
DK3S	Deutschleistung Klasse 3 (Schreiben)	0 – 54 (Summenscore)
BLSF.D	Förderbedarf in Deutsch	0 = *kein Förderbedarf*, 1 = *Förderbedarf*
SES.D	SES[a] der Deutsch-Klasse	ordinal[b]; Kategorien 1 – 4
SART.D	Schulart der Deutsch-Klasse	1 = *Förderschule*, 2 = *Hauptschule*, 3 = *nicht-differenziert*[c], 4 = *Realschule*, 5 = *Gymnasium*
Fachunspezifische Variablen		
SEX	Geschlecht	0 = *männlich*, 1 = *weiblich*
MUSPR	Muttersprache	0 = *nicht-deutsch*, 1 = *deutsch*
WDH	Wiederholer	0 = *nein*, 1 = *ja*

Anmerkungen. Der erste und zweite Teil der Tabelle enthält alle für das Fach Mathematik bzw. Deutsch spezifischen Variablen. Im dritten Teil der Tabelle sind die fachunspezifischen Variablen aufgelistet, die sowohl im Rahmen der Adjustierung im Fach Mathematik als auch im Fach Deutsch verwendet werden.

[a] SES = Sozioökonomischer Status (engl.: socio-economic status).

[b] Kategorisierung der Variable *Bücherfrage* in Quartile.

[c] Die Kategorie *nicht-differenziert* betrifft bspw. Klassen in Gemeinschaftsschulen, in denen noch keine Differenzierung des Bildungsganges erfolgt ist.

kator des sozioökonomischen Status (SES) bzw. der Bildungsnähe erwiesen (vgl. Bos et al., 2003; M. D. Evans, Kelley, Sikora & Treiman, 2010) und wird daher seit dem Schuljahr 2003/2004 auch in dem Fragenkatalog des Kompetenztests verwendet (Nachtigall, Hempel, Jantowski, Kröhne & Müller, 2004). Seit dem Schuljahr 2008/2009 wird die Bücherfrage allerdings nur noch in anonymisierter Form auf Klassenebene erhoben. Das bedeutet, dass für jede Klasse lediglich der Mittelwert des SES über die Schüler dieser Klasse vorliegt. Daher haben alle Schüler einer konkreten Klasse die gleiche Ausprägung auf der Variable SES. Die SES-Werte eines konkreten Schülers im Fach Mathematik (SES.M) im Vergleich zu Deutsch (SES.D) können sich dahingegen jedoch unterscheiden in Abhängigkeit von der jeweiligen Klassenzusammensetzung in beiden Fächern. Schließlich wird auch die Art der Schule (SART), der ein Schüler angehört, erfasst und in der Datenauswertung berücksichtigt. Hierbei wird zwischen *Förderschule*, *Hauptschule*, *Realschule* und *Gymnasium* differenziert. Des Weiteren werden auch sog. *nicht-differenzierte* Bildungsgänge unterschieden, welche an Gesamtschulen vorkommen. Hier werden je nach angestrebtem Abschluss entsprechende Kurse gebildet, wobei in nicht-differenzierten Kursen noch keine Trennung des Bildungsganges erfolgt ist. Die Kategorien der Variable SART bezeichnen somit nicht nur separate Schularten, sondern gleichfalls Bildungsgänge bzw. Kurse innerhalb einer Schule. So kann z. B. ein Schüler einer Gesamtschule im Fach Mathematik einem Hauptschulkurs, im Fach Deutsch hingegen einem Realschulkurs angehören.

Weiterhin wird bei der Berechnung des korrigierten Landesmittelwertes auch das fachspezifische Vorwissen der Schüler berücksichtigt. Das fachspezifische Vorwissen wird über die Testleistung der Schüler in vorangegangenen Kompetenztests operationalisiert. Hierzu werden die Daten aus dem Thüringer Schülerlängsschnitt verwendet, welche eine Zuordnung der Kompetenztestleistungen im Fach Mathematik sowie Deutsch aus früheren Klassenstufen ermöglicht. Wie bereits in Abschnitt 6.1 erläutert lagen im Schuljahr 2009/2010 erstmals Daten aus drei Erhebungswellen vor (vgl. Tabelle 6.1). Somit können im Fach Mathematik und Deutsch sowohl die Leistungsdaten aus Klassenstufe 6 als auch Klassenstufe 3 als Indikator für das Vorwissen in den jeweiligen Fächern im Rahmen der Datenauswertung genutzt werden. Im Kompetenztest zur Erhebung der Mathematikleistung in Klassenstufe 6 (MK6) im Schuljahr 2007/2008 konnten maximal 28 Punkte erreicht werden. Bei dem im gleichen Schuljahr durchgeführten Kompetenztest zur Erfassung der Deutschleistung in Klassenstufe 6 (DK6) konnten maximal 96 Punkte erzielt werden. Bei der Erhebung der Mathematikleistung

mit dem Kompetenztest in Klassenstufe 3 (MK3) im Schuljahr 2007/2008 konnten maximal 24 Punkte erreicht werden. Die Deutschleistung in Klassenstufe 3 wurde in jenem Schuljahr mit dem Kompetenztest Deutsch-Lesen (DK3L) und dem Kompetenztest Deutsch-Schreiben (DK3S) erhoben, in welchen maximal 15 respektive 54 Punkte erreicht werden konnten. Für die nachfolgenden Analysen werden jeweils die Summenscores der Kompetenztests betrachtet. Diese Summenscores der fachspezifischen Kompetenztests sind jedoch zwischen den Schuljahren nicht direkt miteinander vergleichbar, da die Tests auf unterschiedlichen Skalen gemessen wurden. Ein Schüler, der in den Mathematiktests der dritten und sechsten Klasse jeweils 24 Punkte erreichte, weist also in Klassenstufe 6 eine geringere – und nicht etwa konstante – Mathematikleistung auf.

Fachunspezifische Kovariaten

Weitere Schülereigenschaften, die unabhängig vom jeweiligen Fachbereich erfasst wurden, sind das Geschlecht des Schülers (SEX: 0 = *männlich*, 1 = *weiblich*), die Muttersprache (MUSPR: 0 = *nicht-deutsch*, 1 = *deutsch*) sowie die Wiederholung der achten oder einer früheren Klassenstufe (WDH: 0 = *nein*, 1 = *ja*).

6.3 Statistische Methoden

Zur Prüfung der in Kapitel 5 postulierten Hypothesen wurde ein empirischer Modellvergleich verschiedener Adjustierungsmodelle durchgeführt. Im folgenden Abschnitt werde ich das Design dieses Modellvergleichs sowie die zur Prüfung der Hypothesen verwendeten Kriterien darstellen. Anschließend werden die Herausforderungen und Möglichkeiten des Umgangs mit fehlenden Werten (*Missing Data*) diskutiert. Dies bildet die Grundlage bzw. den Ausgangspunkt für den Umgang mit fehlenden Werten im Rahmen der vorliegenden Arbeit. Sämtliche Analysen wurden mit der freien Software R (R Development Core Team, 2010) durchgeführt.

6.3.1 Design des Modellvergleichs

Nachfolgend werden die für den empirischen Modellvergleich verwendeten Modelle dargestellt. Die einzelnen Modelle lassen sich hinsichtlich der gewählten Kovariaten

(Kovariatenselektion) sowie der gewählten Parametrisierung (Modellselektion) unterscheiden. Wie bereits in Kapitel 5 dargestellt, ist die vorliegende empirische Reanalyse als eine Sensitivitätsanalyse aufzufassen. Daher werden im folgenden Abschnitt gleichfalls die Kriterien des Modellvergleich vorgestellt, anhand derer die Sensitivität der Effektschätzungen quantifiziert wird.

Sämtliche Modelle des Modellvergleichs zielen auf die Berechnung eines klassenspezifischen Effekts. Da sich die Reanalyse auf Daten aus dem Projekt *Kompetenztest.de* bezieht, wird die dort verwendete Adjustierungsstrategie (Strategie IV; vgl. Kapitel 4) gleichermaßen Ausgangspunkt dieses Modellvergleichs sein. Aus diesem Grund wird im Anschluss zunächst das schrittweise Vorgehen gemäß Strategie IV zur Berechnung des adjustierten klassenspezifischen Effekts dargestellt.

Das Vorgehen bei der Datenanalyse im Projekt *Kompetenztest.de*

Im Projekt *Kompetenztest.de* wird ein Effektivitätsmaß für den Unterricht einzelner Klassen, d. h. auf Klassenebene, berechnet. Dieser adjustierte klassenspezifische Effekt wird nachfolgend mit $E(\delta_{adj} \mid X{=}x)$ notiert (vgl. Kapitel 3). Zur Quantifizierung von $E(\delta_{adj} \mid X{=}x)$ wird der beobachtete Mittelwert einer Klasse x einem für eine virtuelle Klasse mit gleicher Kovariatenkombination zu erwartenden Wert, dem sog. *korrigierten Landesmittelwert*, gegenübergestellt (Fiege, 2007; Nachtigall & Kröhne, 2003; Nachtigall et al., 2008), d. h.:

$$E(\delta_{adj} \mid X{=}x) \equiv E(Y \mid X{=}x) - E\big[\, E(Y \mid \mathbf{Z}) \mid X{=}x \,\big]. \tag{6.1}$$

Diese Differenz wird im Rahmen der klassenspezifischen Ergebnisrückmeldungen u. a. grafisch veranschaulicht bzw. in grafischer Form zurückgemeldet. Abbildung 6.3 zeigt ein Beispiel einer solchen Darstellung, die einem Ergebnisbericht des Projektes *Kompetenztest.de* zu entnehmen ist (vgl. Nachtigall, Müller & Storbeck, 2010). Wie bereits in Kapitel 3 und 4 dargestellt, bezieht sich die Adjustierung somit auf den zu berechnenden Vergleichswert $E\big[\, E(Y \mid \mathbf{Z}) \mid X{=}x \,\big]$ für die jeweils betrachtete Klasse x.

Die Berechnung bzw. Schätzung der Effekte auf Ebene der einzelnen Klassen, d. h. der Differenz zwischen Klassenmittelwert und adjustiertem Referenzwert einer Klasse, erfolgt schrittweise. Diese drei Analyseschritte umfassen (1) die Berechnung des Klassenmittelwertes, (2) die Berechnung des adjustierten Referenzwertes einer Klasse und schließlich (3) die Berechnung des adjustierten klassenspezifischen Effektmaßes:

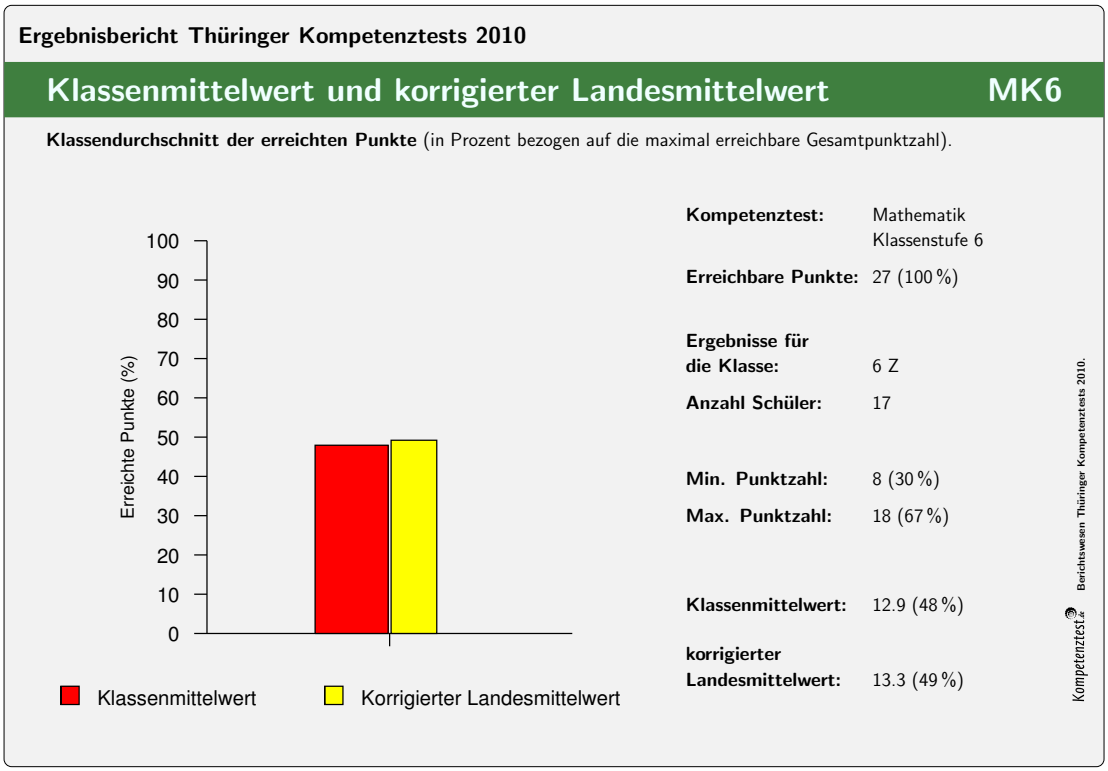

Abbildung 6.3: Rückmeldeformat im Projekt *Kompetenztest.de* (aus Nachtigall et al., 2010). Links: Grafische Darstellung des Klassenmittelwertes und des korrigierten Landesmittelwertes am Beispiel der Mathematikleistung in Klassenstufe 6 (MK6). Rechts: Deskriptive Kennwerte.

(1) *Klassenmittelwert der Testwerte:*
Zunächst werden die Klassenmittelwerte der Testwerte berechnet. Der Klassenmittelwert der Testwertvariablen Y einer Klasse x mit N_x Schülern berechnet sich wie folgt:

$$\begin{aligned} \bar{Y}_x &= \widehat{E}(Y \mid X{=}x) \\ &= \frac{1}{N_x} \sum_{i=1}^{N_x} Y_{ix}. \end{aligned} \tag{6.2}$$

Dabei bezeichnet Y_{ix} den i-ten zu beobachtenden Testwert in der Klasse x. Weiterhin gilt: $\bar{Y}_x = \widehat{E}(Y \mid X{=}x)$, d. h. der Klassenmittelwert $\bar{Y}_x$ ist ein Schätzer für den Populationskennwert $E(Y|X{=}x)$, der im Rahmen des Single-Unit-Trials (vgl. Kapitel 3) definiert ist. Der bedingte Erwartungswert $E(Y \mid X{=}x)$ der Outcome-Variablen Y gegeben $X{=}x$ ist die mit den bedingten Wahrscheinlichkeiten gewichtete Summe ihrer Werte, d. h. $E(Y|X{=}x) = \sum_i y_i \cdot P(Y{=}y_i \,|\, X{=}x)$. Dabei ist Y im Rahmen von Vergleichsarbeiten die Testwertvariable und X die Treatment-Variable mit $x = 1, \ldots, J$ Treatment-Ausprägungen, welche im vorliegenden Kontext die untersuchten Thüringer Klassen der Jahrgangsstufe 8 repräsentieren.

(2) *Adjustierter Referenzwert einer Klasse:*
Der adjustierte Klassenmittelwert bzw. Referenzwert einer Klasse x, der im Projekt *Kompetenztest.de* als korrigierter Landesmittelwert bezeichnet wird, berechnet sich wie folgt: Zunächst wird für jeden Schüler der betrachteten Klasse x der bedingte Erwartungswert $E(Y \mid \boldsymbol{Z}{=}\boldsymbol{z})$ bei gegebener Kovariatenausprägung geschätzt. Dies erfolgt mittels der Regression $E(Y \mid \boldsymbol{Z})$ der Testwertvariablen Y auf den Kovariatenvektor $\boldsymbol{Z}$, wobei $\boldsymbol{Z}$ den Vektor einer Q-dimensionalen Kovariaten $\boldsymbol{Z} = (Z_1, \ldots, Z_Q)$ darstellt. Im Rahmen des Modellvergleichs der vorliegenden Arbeit wurden unterschiedliche Parametrisierungen, d. h. unterschiedliche Modellspezifikationen der Regression $E(Y \mid \boldsymbol{Z})$ sowie verschiedene Sets von Kovariaten gewählt, die im nächsten Abschnitt beschrieben werden. Die mittels der Regression $E(Y \mid \boldsymbol{Z})$ vorhergesagten Werte werden anschließend über die Schüler einer Klasse aggregiert, d. h. es wird der $X{=}x$-bedingte Erwartungswert

$E[\, E(Y \mid \mathbf{Z}) \mid X{=}x \,]$ geschätzt:

$$\begin{aligned} \bar{Y}_{adj;x} &= \widehat{E}[\, \widehat{E}(Y \mid \mathbf{Z}) \mid X{=}x \,] \\ &= \frac{1}{N_x} \cdot \sum_{i=1}^{N_x} \widehat{E}(Y \mid \mathbf{Z}{=}\mathbf{z}_i)\,, \end{aligned} \tag{6.3}$$

wobei $\widehat{E}(Y \mid \mathbf{Z}{=}\mathbf{z}_i)$ die Schätzer der adjustierten individuellen (schülerspezifischen) Testwerte sind. Weiterhin gilt: $\bar{Y}_{adj;x} = \widehat{E}[\, \widehat{E}(Y|\mathbf{Z})|X{=}x \,]$, d. h., der Klassenmittelwert $\bar{Y}_{adj;x}$ ist ein Schätzer für den Populationskennwert $E[\, E(Y \mid \mathbf{Z}) \mid X{=}x \,]$, der im Rahmen des Single-Unit-Trials (vgl. Kapitel 3) definiert ist.

(3) *Adjustierte klassenspezifischer Effekt:*
Schließlich wird das klassenspezifische Effektmaß $E(\delta_{adj} \mid X{=}x)$ geschätzt:

$$\begin{aligned} \bar{\delta}_{adj;x} &= \widehat{E}(Y \mid X{=}x) - \widehat{E}[\, \widehat{E}(Y \mid \mathbf{Z}) \mid X{=}x \,] \\ &= \bar{Y}_x - \bar{Y}_{adj;x}\ . \end{aligned} \tag{6.4}$$

Es gilt: $\bar{\delta}_{adj;x} = \widehat{E}(\delta_{adj} \mid X{=}x)$, d. h. das klassenspezifische Effektmaß $\bar{\delta}_{adj;x}$ ist ein Schätzer für den Populationskennwert $E(\delta_{adj} \mid X{=}x)$, der wiederum im Rahmen des zugrundeliegenden Single-Unit-Trials (vgl. Kapitel 3) definiert ist. Wie bereits ausführlich dargestellt, impliziert dies jedoch nicht ohne weitere Annahmen die kausale Interpretierbarkeit von $E(\delta_{adj} \mid X{=}x)$ oder seines Schätzers $\bar{\delta}_{adj;x}$ (vgl. Kapitel 3, Abschnitt 3.5).

Modelle im Vergleich

Im Folgenden werde ich die Modelle vorstellen, deren Ergebnisse im Kapitel 7 vergleichend betrachtet werden. Diese Modelle unterscheiden sich einerseits hinsichtlich der Kovariatenselektion und andererseits bezüglich der Modellselektion, d. h. hinsichtlich der gewählten Parametrisierung. Insgesamt 14 Modelle wurden analysiert – sowohl im Fachbereich Mathematik als auch Deutsch. Tabelle 6.4 zeigt die Modelle schließlich im Überblick.

Kovariatenselektion: Die Wahl der Kovariaten. Die zu vergleichenden Modelle unterscheiden sich hinsichtlich der enthaltenen Kovariaten, wobei verschiedene Sets

von Kovariaten berücksichtigt werden. Diese Kovariatensets lassen sich in drei Gruppen einteilen: (a) Schülermerkmale (ohne Vorwissen), (b) Schülermerkmale und fachspezifisches Vorwissen sowie (c) Schülermerkmale, fachspezifisches Vorwissen und Klassenkompositionsmerkmale. Je nach Kovariatenset können die Modelle demnach den verschiedenen Klassen von Adjustierungsmodellen zugeordnet werden, die in Kapitel 4 eingeführt wurden: (a) *Contexualized Attainment Modelle* (CAM) ohne Vorwissen (b) *Value-Added Modelle* (VAM), die neben Schülermerkmalen auch das Vorwissen berücksichtigen und schließlich (c) *Contextual Value-Added Modelle* (CVA), die zusätzlich zu Schülermerkmalen und Vorwissen auch Kompositionsmerkmale enthalten. Bei gegebener Modellselektion (d. h. bei konstanter Parametrisierung der Modelle) handelt es sich dabei um *genestete Modelle*: Durch die Hinzunahme weiterer Kovariaten werden die Modelle jeweils erweitert. So ist bspw. das CAM ein (restriktiverer) Spezialfall vom VAM, in dem das fachspezifische Vorwissen als Kovariate *nicht* enthalten ist. Im Falle bedingter regressiver Unabhängkeit der Testwertvariablen Y vom fachspezifischen Vorwissen gegeben der weiteren, in beiden Modellen gewählten Kovariaten werden beide Modelle identische Ergebnisse liefern.

In den verschiedenen Kovariatensets bzw. Modellklassen sind jeweils folgenden Kovariaten enthalten:

(1) *CAM (Modell 1 und Modell 8):*
Zunächst werden lediglich Schülermerkmale – und explizit nicht das Vorwissen der Schüler – berücksichtigt. Im Einzelnen handelt es sich um die folgenden diskreten Kovariaten: Diagnose besonderer Lernschwierigkeiten bzw. sonderpädagogischer Förderbedarf (BLSF), Anzahl der Bücher im Elternhaus (SES), Schulart (SART), Geschlecht (SEX), Muttersprache (MUSPR) und Wiederholer einer Klassenstufe (WDH).

(2) *VAM (Modelle 2 bis 4 und Modelle 9 bis 11):*
Zusätzlich zu den Schülermerkmalen wird nun das fachspezifische Vorwissen in das Kovariatenset aufgenommen. Zunächst wird das fachspezifische Vorwissen der 3. Jahrgangsstufe (Modelle 2 und 9), weiterhin das fachspezifische Vorwissen der 6. Jahrgangsstufe (Modelle 3 und 10) und schließlich das fachspezifische Vorwissen sowohl aus der 3. als auch 6. Jahrgangsstufe (Modelle 4 und 11) in das Modell aufgenommen.

(3) *CVA (Modelle 5 bis 7 und Modelle 12 bis 15):*
Im dritten Kovariatenset werden – neben Schülermerkmalen und fachspezifischen Vorwissen – zusätzlich Klassenkompositionsmerkmale berücksichtigt. Die Klassenkomposition wird operationalisiert mittels des Mittelwertes und der Standardabweichung des fachspezifischen Vorwissens. Dabei wird zunächst das fachspezifische Vorwissen aus Klasse 3 (Modelle 5 und 12), dann aus Klasse 6 (Modelle 6 und 13) und letztlich aus Klasse 3 *und* Klasse 6 (Modelle 7 und 14) für die Berechnung der Klassenkomposition verwendet.

Modellselektion: Die Parametrisierungen der Regression $E(Y \mid \boldsymbol{Z})$. Die für den empirischen Modellvergleich gewählten Modellspezifikationen basieren einerseits auf der vom Projekt *Kompetenztest.de* derzeit verwendeten saturierten Parametrisierung sowie deren Erweiterung als bedingt lineare Parametrisierung bei Hinzunahme eines metrischen Regressors (Modelle 1 bis 7). Andererseits wurde eine weniger komplexe Parametrisierung (lineare Parametrisierung ohne Interaktionen) für den Modellvergleich gewählt (Modelle 8 bis 14). Diese Parametrisierungen werden nachfolgend im Detail beschrieben.

(1) *Saturierte Parametrisierung (Modell 1):*
Bisher wird im Rahmen der Ergebnisauswertung durch das Projekt *Kompetenztest.de* ein saturiertes Modell verwendet, welches für metrische Variablen nicht anwendbar ist. Ein saturiertes Modell erfordert diskrete oder diskretisierte (kategoriale oder künstlich kategorisierte) Variablen. Im Rahmen der Adjustierungsstrategie IVa (vgl. Kapitel 4) des Projektes *Kompetenztest.de*, die in Modell 1 des vorliegenden Modellvergleichs realisiert ist, werden folgende diskrete Kovariaten berücksichtigt: Diagnose besonderer Lernschwierigkeiten bzw. sonderpädagogischer Förderbedarf (BLSF), Anzahl der Bücher im Elternhaus (SES), Schulart (SART), Geschlecht (SEX), Muttersprache (MUSPR) und Wiederholer einer Klassenstufe (WDH). Diese Kovariaten bilden gemeinsam den mehrdimensionalen Regressor $\boldsymbol{Z}$, dessen Werte die Wertekombinationen der genannten Kovariaten sind. Im Fall diskreter Kovariaten nimmt der mehrdimensionale Regressor $\boldsymbol{Z}$ genau m verschiedene Werte an, wobei m die Anzahl aller möglichen Wertekombinationen der mehrdimensionalen Kovariaten $\boldsymbol{Z}$ ist. In diesem Fall kann die Regression $E(Y \mid \boldsymbol{Z})$ als saturiertes Zellenmittelwertemodell para-

metrisiert werden:

$$E(Y \mid \boldsymbol{Z}) \quad = \quad \mu_1 \cdot I_{\boldsymbol{Z}=1} + \ldots + \mu_z \cdot I_{\boldsymbol{Z}=z} + \ldots + \mu_m \cdot I_{\boldsymbol{Z}=m}\,, \tag{6.5}$$

wobei

$$I_{\boldsymbol{Z}=z} = \begin{cases} 1, \text{ falls } \boldsymbol{Z}=z, \\ 0, \text{ andernfalls} \end{cases} \qquad \text{für } z = 1, \ldots, m$$

jeweils die Indikatorvariable ist, die mit ihrem Wert 1 anzeigt, ob $\boldsymbol{Z}$ den Wert z annimmt. Die Anzahl der Parameter μ_z entspricht hier der Anzahl der Wertekombinationen des mehrdimensionalen Regressors $\boldsymbol{Z}$. Des Weiteren ist μ_z der bedingte Erwartungswert $E(Y \mid \boldsymbol{Z}=z)$ der Outcome-Variablen Y gegeben eine bestimmte Kovariatenkombination z. Dieses saturierte Modell ist dadurch gekennzeichnet, dass hier keinerlei Annahmen über die funktionale Form der Abhängigkeit der Zufallsvariablen Y von dem Kovariatenvektor $\boldsymbol{Z}$ gemacht werden, die sich in der empirischen Anwendung als falsch erweisen können.

(2) *Bedingt lineare Parametrisierung (Modelle 2 bis 7):*
Dem Vorteil der saturierten Parametrisierung steht der Nachteil gegenüber, dass sich im Rahmen dieser Vorgehensweise keine kontinuierlichen Kovariaten berücksichtigen lassen. Ein zusätzliches Problem hinsichtlich der Parameterschätzung besteht potenziell darin, dass man stets m Parameter – entsprechend der Anzahl der Wertekombinationen der Kovariaten $\boldsymbol{Z}$ – schätzen muss. Je mehr Kovariaten nun in der empirischen Anwendung berücksichtigt werden, desto größer wird das Problem der Parameterschätzung: Sind bestimmte Zellen in den Daten nicht oder mit sehr wenigen Beobachtungen besetzt, können die entsprechenden Erwartungswerte gar nicht oder nur mit großem Standardfehler geschätzt werden (*Sparseness of Data*-Problem, vgl. Fiege, 2007; Morgan & Harding, 2006). Da zudem die Kategorisierung des fachspezifischen Vorwissens mit einem Informationsverlust verbunden ist, der sich in deutlichem Maße auf die Effektschätzungen auswirken kann (z. B. Sengewald, 2011), wird in der vorliegende Reanalyse der Kompetenztestdaten die metrische Variable $Z_{Vorwissen}$ einbezogen. In diesem Fall bilden der mehrdimensionale Regressor $\boldsymbol{Z}$, dessen Werte die Wertekombinationen der verwendeten diskreten Kovariaten sind, zusammen mit der kontinuierli-

chen Kovariaten $Z_{Vorwissen}$ den mehrdimensionalen Regressor $\boldsymbol{Z}^* = (\boldsymbol{Z}, Z_{Vorwissen})$. Für die Modelle 2 bis 7, die das fachspezifische Vorwissen enthalten, verwende ich daher eine bedingt lineare Parametrisierung, d. h.:

$$E(Y \mid \boldsymbol{Z}^*) = \boldsymbol{\beta}_{01} \cdot \mathbf{I}_{\mathbf{Z}} + \boldsymbol{\beta}_{02} \cdot \mathbf{I}_{\mathbf{Z}} \cdot Z_{Vorwissen}\,, \tag{6.6}$$

wobei

$$\boldsymbol{\beta}_{01} = (\beta_1, \ldots, \beta_m) \tag{6.7}$$

und

$$\boldsymbol{\beta}_{02} = (\beta_{m+1}, \ldots, \beta_{2 \cdot m})\,. \tag{6.8}$$

Dabei ist $\boldsymbol{\beta}_{02}$ eine Funktion von $\boldsymbol{Z}$ und $Z_{Vorwissen}$. Auf diese Weise werden potenzielle Interaktionen zwischen dem mehrdimensionalen, kategorialen Regressor $\boldsymbol{Z}$ und dem metrischen Regressor $Z_{Vorwissen}$ modelliert. Setzt man Gleichung 6.7 und 6.8 in Gleichung 6.6 ein, so ergibt sich:

$$E(Y \mid \boldsymbol{Z}^*) = (\beta_1, \ldots, \beta_m) \cdot \begin{pmatrix} I_{\mathbf{Z}=1} \\ \vdots \\ I_{\mathbf{Z}=m} \end{pmatrix} + (\beta_{m+2}, \ldots, \beta_{2 \cdot m+1}) \cdot \begin{pmatrix} I_{\mathbf{Z}=1} \\ \vdots \\ I_{\mathbf{Z}=m} \end{pmatrix} \cdot Z_{Vorwissen}\,. \tag{6.9}$$

Bei der bedingt linearen Parametrisierung wird die Annahme gemacht, dass die Regression $E(Y \mid \boldsymbol{Z}^*)$ bedingt linear in $Z_{Vorwissen}$ ist. Dabei sind die Parameter des Vektors $\boldsymbol{\beta}_{02}$ in Gleichung 6.6 die *bedingt linearen Regressionskoeffizienten* (vgl. Steyer, 2003). Empirische Befunde stützen die Plausibilität dieser Annahme. So zeigte z. B. Sengewald (2011) im Kontext von Vergleichsarbeiten, dass es keine bedeutsamen Unterschiede bezüglich der klassenspezifischen Effektschätzungen gibt, wenn anstelle von $\boldsymbol{Z}^* = (\boldsymbol{Z}, Z_{Vorwissen})$ der mehrdimensionale Regressor $\boldsymbol{Z}^{**} = (\boldsymbol{Z}, Z_{Vorwissen}, Z^2_{Vorwissen})$ verwendet wird[6]. Im letzteren Fall handelt es sich ebenfalls um eine bedingt lineare Parametrisierung: Die Regression $E(Y \mid \boldsymbol{Z}^{**})$ ist

[6] Neben der linearen und quadratischen Parametrisierung der kontinuierlichen Kovariaten $Z_{Vorwissen}$ wurden auch kubische, quartische und quintische Parametrisierungen in die vergleichende Analyse einbezogen. Diese erbrachten ebenfalls keine bedeutsamen Unterschiede in den Effektschätzungen (vgl. Sengewald, 2011).

bedingt linear in $Z_{Vorwissen}$ und $Z^2_{Vorwissen}$.

(3) *Lineare Parametrisierung (Modelle 8 bis 14):*
Als alternative, sparsamere Modellspezifikation verwende ich eine lineare Parametrisierung (Modelle 8 bis 14). Eine solche lineare Parametrisierung wurde bspw. im Rahmen der Ergebnisauswertung der nationalen Erweiterung von PISA-2000 in der Bundesrepublik Deutschland angewendet (Watermann, Stanat, Kunter, Klieme & Baumert, 2003; Watermann & Stanat, 2004). Die Regression $E(Y \mid \boldsymbol{Z})$ wird mittels einer linearen Funktion des mehrdimensionalen Regressors $\boldsymbol{Z}$ parametrisiert:

$$E(Y \mid \boldsymbol{Z}) = \alpha_0 + \alpha_1 \cdot Z_1 + \cdots + \alpha_Q \cdot Z_Q \,, \tag{6.10}$$

wobei $\boldsymbol{Z}$ der Vektor einer Q-dimensionalen Kovariaten $\boldsymbol{Z} = (Z_1, \ldots, Z_Q)$ ist. Potenzielle Interaktionen zwischen den Kovariaten werden hier nicht modelliert. Bei der Berechnung des $\boldsymbol{Z}{=}z$-bedingten Erwartungswertes $E(Y \mid \boldsymbol{Z}{=}z)$ können nicht nur diskrete, sondern gleichfalls kontinuierliche Kovariaten einbezogen werden. Jedoch fließen hier bestimmte Annahmen ein, die in empirischen Anwendungen falsch sein können: Zum einen wird von einem linearen Zusammenhang zwischen der Outcome-Variable Y und dem Kovariatenvektor $\boldsymbol{Z}$ ausgegangen und zum anderen werden potenzielle Interaktionen zwischen den Kovariaten nicht berücksichtigt.

Der empirische Modellvergleich umfasst somit insgesamt 14 Modelle, die sowohl für den Fachbereich Mathematik (MK8) als auch Deutsch (DK8) berechnet werden. Tabelle 6.4 zeigt die zu vergleichenden Modelle in der Übersicht.

Tabelle 6.4: Modellvergleich

Modell-selektion	Kovariatenselektion						
	CAM	VAM			CVA		
	Schüler-merkmale	+ fachspezif. Vorwissen (K3)	+ fachspezif. Vorwissen (K6)	+ fachspezif. Vorwissen (K3 & K6)	+ Klassen-kompositions-merkmale[a] (K3)	+ Klassen-kompositions-merkmale[a] (K6)	+ Klassen-kompositions-merkmale[a] (K3 & K6)
bedingt lineare Parametrisierung (inkl. Interaktionen)	Modell 1[b]	Modell 2	Modell 3	Modell 4	Modell 5	Modell 6	Modell 7
lineare Parametrisierung (ohne Interaktionen)	Modell 8	Modell 9	Modell 10	Modell 11	Modell 12	Modell 13	Modell 14

Anmerkungen. CAM = Contextualized Attainment Model, VAM = Value-Added Model, CVA = Contextual Value-Added Model, K3 = Kompetenztest Klasse 3, K6 = Kompetenztest Klasse 6.
[a] Mittelwert und Standardabweichung des fachspezifischen Vorwissens aus Klasse 3, Klasse 6 bzw. Klasse 3 & 6.
[b] Saturiertes Zellenmittelwertemodell.

Die einzelnen Spalten der Tabelle 6.4 kennzeichnen die jeweilige Selektion der Kovariaten. Modelle in der gleichen Spalte enthalten das gleiche Set von Kovariaten. Modelle, die in der Tabelle weiter rechts angeordnet sind, enthalten jeweils zusätzliche Kovariaten. Somit lassen sich verschiedene Modelle ineinander überführen[7]: So ist bspw. Modell 1 ein Spezialfall von Modell 2, in dem das fachspezifische Vorwissen aus Klassenstufe 3 als Kovariate *nicht* enthalten ist. Im Falle bedingter regressiver Unabhängkeit der Testwertvariablen Y vom fachspezifischen Vorwissen aus Klassenstufe 3 gegeben der weiteren, in beiden Modellen gewählten Kovariaten werden beide Modelle identische Ergebnisse liefern.

Die Modelle in den beiden Zeilen von Tabelle 6.4 unterscheiden sich hinsichtlich der Parametrisierung des zur Berechnung der adjustierten klassenspezifischen Effekte gewählten Modells. Die erste Zeile enthält die Modelle mit saturierter bzw. bedingt linearer Parametrisierung (Modelle 1 bis 7). Die zweite Zeile enthält die Modelle mit linearer Parametrisierung, bei denen keine Interaktionen zwischen den Variablen modelliert werden (Modelle 8 bis 14).

Kriterien des Modellvergleichs

Für den Vergleich der aus den 14 Modellen resultierenden adjustierten Effektschätzungen – sowohl für das Fach Mathematik als auch das Fach Deutsch – werden im Ergebnisteil der vorliegenden Arbeit (Kapitel 7) folgende Kriterien betrachtet: Zunächst werden die Ergebnisse der einzelnen Modelle in sog. *Caterpillar-Plots* dargestellt. Die Caterpillar-Plots enthalten neben den Punktschätzern der adjustierten klassenspezifischen Effekte auch die zugehörigen Standardfehler. Zudem veranschaulichen Caterpillar-Plots die Rangordnung der Klassen gemäß dem Ausmaß des adjustierten klassenspezifischen Effekts. Weiterhin werden die modellspezifischen Determinationskoeffizienten $R^2_{Y|Z}$ – als Maß der Varianzaufklärung – vergleichend analysiert sowie die Korrelation der aus den Modellen resultierenden adjustierten Effektschätzungen der einzelnen Klassen (vgl. z. B. Braun & Wainer, 2007) ausgewertet. Schließlich werden die Veränderungen der adjustierten klassenspezifischen Effektschätzungen beim Wechsel der Adjustierungsstrategie (vgl. z. B. Briggs & Domingue, 2011) betrachtet. Die Verän-

[7] Die Modelle innerhalb einer Zeile sind genestet. Einzige Ausnahme sind die Modelle, die das fachspezifische Vorwissen aus jeweils unterschiedlichen Jahrgangstufen enthalten. So ist bspw. Modell 2 kein Spezialfall von Modell 3. Die gleiche Einschränkung trifft zu auf die Modelle 5 und 6, 9 und 10 sowie 12 und 13.

derungen der Effektschätzungen beim paarweisen Modellvergleich werden einerseits mittels modifizierter Caterpillar-Plots und andererseits anhand von Kreuztabellen des Quintil-Rankings der Effektschätzungen veranschaulicht. Diese Kriterien dienen jeweils als Maß für die Sensitivität der adjustierten Effektschätzungen gegenüber dem Wechsel des Adjustierungsmodells.

6.3.2 Umgang mit fehlenden Werten

Fehlende Werte (*Missing Data*) stellen im Rahmen der Auswertung empirischer Datensätze häufig ein ernstzunehmendes Problem dar. So können fehlende Werte zu verfälschten Parameterschätzungen führen. Zudem benötigen viele Standardschätzverfahren vollständige Daten, so dass Fälle mit fehlenden Werten aus der Analyse ausgeschlossen werden. Der daraus resultierende Informationsverlust gefährdet die Repräsentativität der Stichprobe sowie die Generalisierbarkeit der Ergebnisse (vgl. Allison, 2001; Graham, 2009; Little & Rubin, 2002; Lüdtke, Robitzsch, Trautwein & Köller, 2007; Peugh & Enders, 2004; Rubin, 1976; Schafer & Graham, 2002; West, 2001).

Dem Problem fehlender Werte wird in der empirischen Bildungsforschung oftmals noch recht wenig Beachtung geschenkt. So beschreiben Peugh und Enders (2004) in ihrem Review über den Umgang mit fehlenden Werten in wissenschaftlichen Publikationen „ … missing data as a ‘dirty little secret’ of educational research“ (S. 540).

Klassifikation fehlender Werte

Für die Auswahl einer geeigneten Methode zum Umgang mit fehlenden Werten ist entscheidend, wie das Fehlen von Werten mit den betrachteten Variablen zusammenhängt. Rubin (1976) beschreibt drei Arten von Missing Data, die den Ausfallprozess (sog. *missing data mechanism*) kennzeichnen. Zum Zweck der Definition dieser drei Arten des Fehlens von Werten betrachte ich – in Anlehnung an Rose (2013) – im Folgenden die Zufallsvariablen Y, Z und R[8]. Dabei sei Y eine Variable, bei der fehlende Werte auftreten können und Z sei eine vollständig beobachtbare Kovariate. Die Variable R sei eine Indikatorvariable für die Beobachtbarkeit der Variablen Y (sog. *response indicator*

[8] Alle drei betrachteten Zufallsvariablen Y, Z und R haben dabei eine gemeinsame Verteilung auf dem zugrunde liegenden Wahrscheinlichkeitsraum $(\Omega, \mathfrak{A}, P)$.

oder Antwort-Indikator), wobei:

$$R = \begin{cases} 1, \text{ falls } Y \text{ beobachtbar ist,} \\ 0, \text{ andernfalls.} \end{cases} \tag{6.11}$$

Nimmt der Response-Indikator R den Wert 1 an, so ist Y beobachtbar. Nimmt der Response-Indikator R hingegen den Wert 0 an, so ist Y nicht beobachtbar. Die drei Arten von Missings nach Rubin (1976) charakterisieren, wie die betrachteten Variablen Y, Z und R stochastisch zusammenhängen. Dabei werden *vollständig zufällig*, *zufällig* und *nicht zufällig* fehlende Werte unterschieden:

(1) Ist die Wahrscheinlichkeit des Auftretens fehlender Werte bezüglich der Variablen Y *vollständig zufällig* (MCAR, *missing completely at random*), so gilt:

$$P(R = 1 \mid Y, Z) = P(R = 1). \tag{6.12}$$

Es gibt also keinen stochastischen Zusammenhang zwischen der Wahrscheinlichkeit fehlender Werte bezüglich Y und der Variablen Y selbst oder der Kovariaten Z. Gilt dies für alle betrachteten Variablen, dann können die Lücken im Datensatz, d. h. die fehlenden Werte, als eine Zufallsstichprobe aus dem Datensatz angesehen werden (Lüdtke et al., 2007). Ein Beispiel soll dies verdeutlichen: Personen, die an einer Fragebogenuntersuchung zur Lebenszufriedenheit (repräsentiert durch die Variablen Y) teilnehmen, sollen außerdem ihre Motivation zur Teilnahme an der Untersuchung (repräsentiert durch die Variablen Z) einschätzen. MCAR liegt dann vor, wenn die Wahrscheinlichkeit der Beantwortung bzw. Nichtbeantwortung der Items zur Lebenszufriedenheit weder von der Höhe der Lebenszufriedenheit noch der Motivation abhängt.

(2) Ist die Wahrscheinlichkeit fehlender Werte bezüglich Y *zufällig* (MAR, *missing at random*), bedeutet dies, dass die Wahrscheinlichkeit des Auftretens fehlender Werte zwar von der Kovariaten Z abhängt, nicht jedoch von der Variablen Y:

$$P(R = 1 \mid Y, Z) = P(R = 1 \mid Z) \quad \text{und} \quad P(R = 1 \mid Y) \neq P(R = 1). \tag{6.13}$$

MAR ist somit eine Form der *bedingten* stochastischen Unabhängigkeit, d. h. gegeben der Kovariaten Z ist die Wahrscheinlichkeit fehlender Werte unabhängig

von der Variablen Y selbst. Bezogen auf unser Beispiel liegt MAR somit dann vor, wenn vor allem Personen mit geringer Motivation Fragen zur Lebenszufriedenheit nicht beantworten. Anders formuliert gibt es hier also einen systematischen Motivationsunterschied zwischen Personen mit Missings und Personen ohne Missings auf der Variable Lebenszufriedenheit. Personen mit gleicher Motivation haben jedoch jeweils die gleiche Wahrscheinlichkeit fehlende Werte hinsichtlich der Lebenszufriedenheit zu zeigen, unabhängig davon, wie hoch bzw. niedrig diese ist.

(3) Bei *nicht zufällig* fehlenden Werten (MNAR, *missing not at random*) hängt die Auftretenswahrscheinlichkeit fehlender Werte bezüglich der Variablen Y von der Variable selbst ab, und zwar auch nachdem für die Kovariate Z kontrolliert wurde:

$$P(R = 1 \mid Y, Z) \neq P(R = 1 \mid Z). \tag{6.14}$$

In unserem Beispiel liegt MNAR dann vor, wenn – auch nach Berücksichtigung der Teilnahmemotivation – die Wahrscheinlichkeit fehlender Werte von der Höhe der Lebenszufriedenheit abhängen.

Diese drei Arten von Missings[9] charakterisieren, wie die betrachteten Variablen und die Auftretenswahrscheinlichkeit fehlender Werte stochastisch zusammenhängen. Weisen die Variablen diese Zusammenhänge auf, dann resultieren bestimmte Muster fehlender Werte in den Daten. Die drei Kategorien – MCAR, MAR und MNAR – schließen sich nicht gegenseitig aus, d. h. in einem konkreten Datensatz können alle drei Arten fehlender Werte vorliegen (Yuan & Bentler, 2000). Sobald jedoch für eine der zu analysierenden Variablen bspw. MAR vorliegt, spricht man nicht mehr von MCAR für den Datensatz.

Die drei Arten fehlender Werte bilden jedoch nicht nur eine Systematik des Missing-Mechanismus, sondern repräsentieren gleichzeitig auch die Annahmen für die verschiedenen Methoden zum Umgang mit fehlenden Werten. Sind diese Annahmen in einer konkreten empirischen Anwendung erfüllt, so führen die entsprechenden Methoden zu unverfälschten Parameterschätzungen.

Ob die Annahme MCAR verletzt ist, d. h., ob ein Zusammenhang zwischen der Wahrscheinlichkeit von Missings und den übrigen Informationen im Datensatz besteht,

[9] Siehe auch Rose (2013) für eine allgemeinere Definition der drei Arten fehlender Werte.

lässt sich bspw. mittels grafischer Darstellungen der Struktur fehlender Werte untersuchen. Diese werde ich in Kapitel 7 (vgl. Abschnitt 7.1.2) am Beispiel der vorliegenden Daten näher erläutern. Darüber hinaus existiert ein multivariater Test (*Little's MCAR-Test*; Little, 1988), welcher prüft, ob die Annahme MCAR verletzt ist. Die Nullhypothese dieses Tests lautet, dass die Wahrscheinlichkeit des Auftretens von Missings bei einer Variable nicht von anderen Variablen innerhalb des Datensatzes abhängt. Einschränkend muss hier jedoch erwähnt werden, dass Little's MCAR-Test lediglich eine Falsifikation der MCAR-Annahme ermöglicht. Mit anderen Worten: Wird der Test nicht signifikant, kann nicht ausgeschlossen werden, dass die Wahrscheinlichkeit des Fehlens bezüglich einer Variable mit der Variable selbst zusammenhängt und somit MNAR vorliegt. Im Gegensatz zu MCAR sind die Annahmen MAR und MNAR einer empirischen Prüfung – im Sinne einer möglichen Falsifikation – nicht zugänglich. Die Ursache hierfür liegt darin, dass man die dazu erforderlichen Ausprägungen der fehlenden Werte des Datensatzes nicht kennt. Eine Entscheidung, ob die Annahme MAR oder aber MNAR für einen konkreten Datensatz plausibel ist, basiert somit in erster Linie auf inhaltlichem Wissen über den Forschungsgegenstand.

Verfahren zum Umgang mit fehlenden Werten

Eine gebräuchliche Ad-hoc-Lösung mit fehlenden Daten umzugehen, ist der Ausschluss aller Fälle, für die auf mindestens einer der zu analysierenden Variablen ein fehlender Wert vorliegt (fallweiser Ausschluss oder *listwise deletion*). Die Nachteile eines solchen Vorgehens sind in der Literatur häufig diskutiert (vgl. Rubin, 1976; Lüdtke et al., 2007). Der fallweise Ausschluss fehlender Werte führt u. U. nicht nur zu einer erheblichen Reduktion der für die Analyse genutzten Daten und damit zu einem Effizienzverlust im Rahmen der Parameterschätzung. Außerdem kann der fallweise Ausschluss zu stark verzerrten Parameterschätzungen führen, wenn der Ausfallprozess nicht vollständig zufällig, d. h. nicht MCAR, ist. Die Ursache für diesen Bias liegt in den systematischen Unterschieden zwischen Fällen mit vollständigen Daten und Fällen mit fehlenden Werten. Schließlich gibt es einen weiteren entscheidenden Nachteil von *listwise deletion*, der insbesondere im Rahmen der vorliegenden Analysestrategie zutrifft: Hier sollen die Ergebnisse verschiedener Modelle miteinander verglichen werden (vgl. 6.3.1). Dieser Modellvergleich soll einen Rückschluss auf die Auswirkung der verschiedenen Modellspezifikationen, die auch die Auswahl verschiedener Kovariatensets einschließt,

ermöglichen. Durch den fallweisen Ausschluss fehlender Werte besteht jedoch die Gefahr, dass die verschiedenen Modelle auf Basis unterschiedlicher Stichproben berechnet werden. Dadurch können Unterschiede in den Ergebnissen nicht mehr eindeutig auf die systematischen Variationen in den Modellspezifikationen attribuiert werden, sondern resultieren möglicherweise aus den Unterschieden in der Datenbasis (oder auch der Interaktion aus beidem).

Neben *listwise deletion* werden in der Literatur eine Vielzahl weiterer Verfahren zum Umgang mit fehlenden Werten diskutiert, die sich in den vergangenen 30 Jahren beträchtlich weiterentwickelt haben (vgl. West, 2001). Gebräuchlich ist dabei eine Klassifikation in traditionelle Methoden (wie bspw. fallweiser oder auch paarweise Ausschluss) und moderne Ansätze zum Umgang mit Missings. Letztere wiederum umfassen neben den modellbasierten Verfahren – wie bspw. die *Full Information Maximum Likelihood* (FIML)-Methode im Kontext von Strukturgleichungsmodellen – auch die imputationsbasierten Verfahren, im Rahmen derer die fehlenden Werte im Datensatz durch geeignete Schätzungen ersetzt werden. Hier ist auch die multiple Imputation nach Rubin (1987) einzuordnen. Bei der multiplen Imputation werden die fehlenden Werte nicht durch einen Wert, sondern durch mehrere plausible Werte (sog. *plausible values*) ersetzt, die üblicherweise jeweils unterschiedliche Ausprägungen aufweisen. Hierbei wird die Annahme gemacht, dass die fehlenden Werte MAR sind. Jedoch zeigen verschiedene Simulationsstudien, dass dieses Verfahren bspw. dem fallweisen Ausschluss auch unter MNAR überlegen ist (vgl. Lüdtke et al., 2007). Als Resultat einer multiplen Imputation erhält man somit mehrere vollständige Datensätze, die unabhängig voneinander analysiert werden. Die Ergebnisse dieser Analysen werden dann nach den Formeln von Rubin (1987) kombiniert. Durch die multiple Imputation wird somit gleichzeitig der Unsicherheit bei der Ersetzung der fehlenden Werte Rechnung getragen, da auch die Variabilität der plausiblen Werte im Rahmen der Kombination der Analyseergebnisse berücksichtigt wird.

Im Rahmen der Auswertung des verwendeten Datensatzes wurde zunächst eine Analyse der Struktur fehlender Werte vorgenommen. Basierend auf den Ergebnissen dieser Analyse wurde schließlich eine geeignete Methode zum Umgang mit potenziell fehlenden Werten im Datensatz ausgewählt.

6.4 Zusammenfassung

Zur Prüfung der in der vorliegenden Arbeit postulierten Hypothesen wurde ein empirischer Modellvergleich durchgeführt. Die zu diesem Zweck verwendeten Daten entstammen den Thüringer Kompetenztests des Schuljahres 2009/2010, in dem erstmals eine längsschnittliche Verknüpfung der Schulleistungsdaten möglich war.

Im vorliegende Kapitel wurden die Erhebungsinstrumente sowie die im Rahmen der statistischen Modelle verwendeten Variablen vorgestellt. Weiterhin wurden die Modelle beschrieben, welche sowohl für den Fachbereich Mathematik als auch Deutsch angewendet wurden. Die insgesamt 14 Modelle unterscheiden sich hinsichtlich der Kovariaten- und Modellselektion. Diese lassen sich aus der Systematik von Adjustierungsstrategien ableiten, welche im Kapitel 4 erarbeitet wurde. Schließlich wurde die Problematik fehlender Werte thematisiert, die gleichfalls Ausgangspunkt der Analyse der Missing-Struktur im verwendeten Datensatz bildete. Basierend auf den Ergebnissen dieser Analyse wurde schließlich eine adäquate Methode zum Umgang mit potenziell fehlenden Werten im Datensatz gewählt.

Die Ergebnisse einer deskriptiven Analyse der verwendeten Daten und der Analyse der Struktur fehlender Werte sind Inhalt des nachfolgenden Kapitels. Im Zentrum des folgenden Kapitels stehen die Ergebnisse des empirischen Modellvergleichs.

Essentially, all models are wrong, but some are useful.

BOX & DRAPER (1986)

7 Ergebnisse: Empirische Befunde aus dem Modellvergleich

Die im vorigen Kapitel erläuterten Modelle wurden auf Daten aus den Thüringer Kompetenztests im Fachbereich Mathematik und Deutsch angewendet. Im folgenden Kapitel werden zunächst die Ergebnisse einer deskriptiven Analyse der verwendeten Daten sowie einer Analyse der Struktur fehlender Werte dargelegt. Anschließend werden die Ergebnisse des empirischen Modellvergleichs dargestellt.

7.1 Deskriptive Analysen

Der Datensatz wurde mir vom Projekt *Kompetenztest.de* zur Verfügung gestellt. Der Datensatz genügt datenschutzrechtlichen Anforderungen, da er keinerlei Information beinhaltet, mittels derer einzelne Thüringer Schulen, Klassen oder Schüler identifiziert werden können. Zwar ist eine Unterscheidung der Schul- und Klassenzugehörigkeit eines Schülers möglich, jedoch erfolgt diese über anonymisierte Codes, welche die Schul- bzw. Klassenzugehörigkeit indizieren[1].

Der Datensatz umfasst die Leistungsdaten von $N = 13\,257$ Thüringer Schülern. Dies sind alle Schüler, die im Schuljahr 2009/2010 an den Kompetenztests im Fach Mathematik und Deutsch der Klassenstufe 8 teilgenommen haben (vgl. Nachtigall, 2010). Des Weiteren beinhaltet der Datensatz die zusammengeführten Leistungsdaten der Schüler aus den verschiedenen Erhebungszeitpunkten (vgl. Tabelle 6.1). Im Rahmen der Datenaufbereitung wurden alle Fälle gelöscht (fallweiser Ausschluss), bei denen entweder keine eindeutige Zuordnung zu einer Schulart (68 Fälle bzw. 0.5%) oder zu einer Klasse (365 Fälle bzw. 2.8%) möglich war. Da die zu berechnenden adjustierten Effektmaße auf Klassenebene ausgewertet werden und statistische Kennwerte – wie das

[1] Die Kodierung der Schulzugehörigkeit erfolgt über die Variable *schule.id* und die Kodierung der Klassenzugehörigkeit erfolgt über die Variable *klasse.id*.

arithmetische Mittel – bei sehr kleinen Fallzahlen wenig aussagekräftig sind, wurden in den Analysen nur solche Klassen berücksichtigt, die aus mindestens fünf Schülern zusammengesetzt sind. Es wurden daher alle Fälle aus der Analyse ausgeschlossen, die sich in einem Klassenverband mit vier oder weniger Schülern[2] befanden (116 Fälle bzw. 0.9%). Nach Ausschluss dieser Fälle reduzierte sich der Datensatz somit auf N = 12 708 Fälle.

7.1.1 Deskriptive Statistiken

Die im Datensatz verbleibenden N = 12 708 Schüler lassen sich N = 365 verschiedenen Thüringer Schulen zuordnen. Tabelle 7.1 enthält die deskriptiven Statistiken für die in Abschnitt 6.2 sowie in Tabelle 6.3 beschriebenen Variablen. Diese charakterisieren die Thüringer Schüler der Klassenstufe 8, die im Schuljahr 2009/2010 an den Kompetenztests Mathematik und Deutsch teilgenommen haben.

Variablen im Fach Mathematik

Im Schuljahr 2009/2010 lag die durchschnittliche Testleistung Thüringer Schüler der Klassenstufe 8 im Kompetenztest Mathematik (MK8) bei M = 17.21 Punkten (SD = 7.22), wobei mindestens eines und maximal sämtliche der 35 Items gelöst wurden. Im Schuljahr 2007/2008 erreichten diese Schüler im Kompetenztest Mathematik der Klassenstufe 6 (MK6) durchschnittlich M = 16.57 Punkte (SD = 5.32). Von den 28 zu lösenden Teilaufgaben wurden minimal keine und maximal alle gelöst. Schließlich wurden im Kompetenztest Mathematik der Klassenstufe 3 (MK3) im Schuljahr 2004/2005 durchschnittlich M = 15.04 von 24 möglichen Punkten erreicht (SD = 4.57), wobei auch hier minimal kein Item und maximal alle Items gelöst wurden.

Der sonderpädagogische Förderbedarf im Fach Mathematik wird durch die Indikatorvariable BLSF.M abgebildet (0 = *kein Förderbedarf*, 1 = *Förderbedarf*; vgl. Tabelle 6.3). Das arithmetische Mittel von Indikatorvariablen mit den Werten 0 und 1 ist gleich der relativen Häufigkeit der Antwortkategorie 1. Demnach zeigten im Schuljahr 2009/2010 rund 5% der Schüler im Fach Mathematik Lernschwierigkeit bzw. sonderpädagogischen Förderbedarf (Tabelle 7.1).

[2] Die durchschnittliche Klassengröße betrug N = 17 Schüler. Alle Klassen, die nur ein Viertel (oder weniger) der durchschnittlichen Klassengröße aufwiesen, wurden aus der Analyse ausgeschlossen.

Tabelle 7.1: Deskriptive Statistiken

Variable	*M*	*SD*	*Min*	*Max*	n^a
Variablen im Fach Mathematik					
MK8	17.21	7.22	1	35	11 590
MK6	16.57	5.32	0	28	9 941
MK3	15.04	4.57	0	24	5 176
BLSF.M	0.05	0.22	0	1	12 629
SES.M	2.73	1.14	1	4	12 658
SART.M[b]	5[c]	—	1	5	12 708
Variablen im Fach Deutsch					
DK8	42.25	11.31	2	68	11 600
DK6	63.76	15.44	0	95	9 951
DK3L	9.07	2.97	0	15	5 161
DK3S	39.77	9.89	0	54	5 198
BLSF.D	0.06	0.23	0	1	12 660
SES.D	2.72	1.12	1	4	12 668
SART.D	5[c]	—	1	5	12 708
Fachunspezifische Variablen					
SEX	0.49	0.50	0	1	12 708
MUSPR	0.97	0.16	0	1	12 626
WDH	0.07	0.25	0	1	12 629

Anmerkungen. Der erste und zweite Teil der Tabelle enthält alle für das Fach Mathematik bzw. Deutsch spezifischen Variablen. Im dritten Teil der Tabelle sind die fachunspezifischen Variablen aufgelistet, die sowohl im Rahmen der Adjustierung im Fach Mathematik als auch im Fach Deutsch verwendet werden.
[a] Anzahl gültiger Werte nach fallweisen Ausschluss fehlender Werte pro Variable.
[b] Schulart ist ein fünfstufiges Merkmal mit folgenden Ausprägungen: 1 = *Förderschule*, 2 = *Hauptschule*, 3 = *nicht-differenziert*, 4 = *Realschule*, 5 = *Gymnasium*.
[c] Da es sich bei der Schulart um ein nominalskaliertes Merkmal handelt, wird hier der Modus anstatt des arithmetischen Mittels als Maß der zentralen Tendenz angegeben.

Der durchschnittliche sozioökonomische Status der Mathematik-Klassen (SES.M) über alle Schüler des Schuljahres 2009/2010 betrug $M = 2.73$ ($SD = 1.14$). Wie in Kapitel 6 dargestellt, kann der SES einer Klasse einen Wert zwischen 1 und 4 annehmen, wobei 1 die niedrigste Stufe und 4 die höchste Stufe des SES indiziert. Im Fach Mathematik gehörten insgesamt 2 683 Schüler (21%) einer Klasse der Stufe 1 des SES an, 2 320 Schüler (19%) einer Klasse der Stufe 2, 3 341 Schüler (26%) einer Klasse der Stufe 3 und schließlich 4 314 Schüler (34%) einer Klasse der Stufe 4. Lediglich 278 Schüler (2%) in Klassenstufe 8 besuchten eine Förderschule (Kategorie 1 der Variable SART.M) und 999 Schüler (8%) besuchten eine Hauptschule (Kategorie 2). Kategorie 3 (*nicht-differenziert*) beschreibt Klassen in Gemeinschaftsschulen, in denen noch keine Differenzierung des Bildungsganges erfolgt ist, der auf einen bestimmten Schulabschluss (Hauptschulabschluss, Realschulabschluss oder allgemeine Hochschulreife) hinausläuft. Dieser Kategorie lassen sich 1 495 Schüler (12%) zuordnen. 4 220 Schüler (33%) besuchten eine Realschulklasse und der Großteil der Thüringer Schüler (5 716 Schüler bzw. 45%) besuchten ein Gymnasium.

Variablen im Fach Deutsch

Im Kompetenztest Deutsch des Schuljahres 2009/2010 lag die durchschnittliche Testleistung Thüringer Schüler der Klassenstufe 8 (DK8) bei $M = 42.25$ von insgesamt 69 erreichbaren Punkten ($SD = 11.31$). Hier wurden mindestens zwei und maximal 68 Punkte erreicht. Im Schuljahr 2007/2008 erreichten diese Schüler im Kompetenztest Deutsch der Klassenstufe 6 (DK6) durchschnittlich $M = 63.76$ Punkte ($SD = 15.44$). Von den zu lösenden Teilaufgaben wurden minimal keine und maximal alle gelöst. Im Schuljahr 2004/2005 wurden im Kompetenztest Deutsch-Lesen der Klassenstufe 3 (DK3L) durchschnittlich $M = 9.07$ von 15 möglichen Punkten erreicht ($SD = 2.97$). Schließlich wurden im Kompetenztest Deutsch-Schreiben der Klassenstufe 3 (DK3S) durchschnittlich $M = 39.77$ von 54 möglichen Punkten erreicht ($SD = 9.89$). Sowohl bei DK3L also auch bei DK3S wurde minimal kein Item und maximal alle Items gelöst.

Ähnlich zu den Ergebnissen der Mathematik-Klassen betrug der durchschnittliche sozioökonomische Status der Deutsch-Klassen (SES.D) $M = 2.72$ ($SD = 1.12$). Auch bezüglich der Häufigkeitsverteilung der Schüler auf die vier Kategorien des SES zeigt sich ein vergleichbares Bild zu den Ergebnissen der Mathematik-Klassen: Im Fach Deutsch gehörten insgesamt 2 514 Schüler (20%) einer Klasse der niedrigsten Stufe

(Stufe 1) des SES an, 2 695 Schüler (21%) einer Klasse der Stufe 2, 3 218 Schüler (26%) einer Klasse der Stufe 3 und 4 241 Schüler (33%) einer Klasse der Stufe 4. Keine Unterschiede zwischen den Fächern Mathematik und Deutsch existieren bezüglich der Zuordnung der Thüringer Schüler in Klassenstufe 8 zu den beiden Schularten Förderschule (278 Schüler bzw. 2%) und Gymnasium (5 716 Schüler bzw. 45%). Für die restlichen Kategorien der Variable SART.D gibt es marginale Unterschiede bezüglich der Häufigkeitsverteilung im Vergleich zu denen in Mathematik: 717 Schüler (6%) besuchten eine Hauptschule (Kategorie 2), 2 644 Schüler (21%) lassen sich Kategorie 3 (*nicht-differenziert*) zuordnen und 3 353 Schüler (26%) einer Realschulklasse (Kategorie 4). Diese Unterschiede liegen darin begründet, dass die Kategorien *Hauptschule*, *nicht-differenziert* und *Realschule* nicht nur separate Schularten angeben, sondern auch Bildungsgänge bzw. Kurse innerhalb einer Schule bezeichnen. So kann z. B. ein Schüler einer Gesamtschule im Fach Mathematik einem Hauptschulkurs angehören, im Fach Deutsch hingegen einem Realschulkurs (vgl. Kapitel 6).

Fachunspezifische Variablen

Bei den fachunspezifischen Variablen Geschlecht (SEX), Muttersprache (MUSPR) und Wiederholung einer Klassenstufe (WDH) handelt es sich – wie auch bei der Variable sonderpädagogischer Förderbedarf (BLSF) – um Indikatorvariablen[3] mit den Werten 0 und 1. Somit ist das arithmetische Mittel dieser Variablen jeweils identisch mit der relativen Häufigkeit der Antwortkategorie 1. Demnach waren im Schuljahr 2009/2010 etwa die Hälfte (49%) der Thüringer Schüler in Klasse 8 weiblich und 97% Prozent der Schüler sprachen deutsch als Muttersprache. Des Weiteren haben 7% der Schüler die achte oder eine frühere Klassenstufe wiederholt.

In der Tabelle 7.1 sind neben den deskriptiven Statistiken der einzelnen Variablen in der letzten Spalte auch die Anzahl *n* gültiger Werte nach fallweisen Ausschluss fehlender Werte pro Variable angegeben. Dabei wird bereits hier deutlich, dass ein größerer Datenausfall auf Variablen aus zurückliegenden Erhebungsjahren – wie bspw. MK6 oder MK3 – zu verzeichnen ist. Nachfolgend soll die Struktur fehlender Werte analysiert werden, welche die Grundlage für die Auswahl eines geeigneten Verfahrens zum Umgang mit den Missings darstellt.

[3] Die entsprechenden Wertelabel der Variablen SEX, MUSPR und WDH finden sich in Tabelle 6.3.

7.1.2 Struktur fehlender Werte

Zum Zweck der Analyse der Missing-Struktur, d. h. der Struktur fehlender Werte im vorliegenden Datensatz, eignen sich insbesondere grafische Darstellungen (vgl. Templ & Alfons, 2009, November). Zur Visualisierung der fehlenden Werte wurde das R-Packet `VIM` (Templ, Alfons & Kowarik, 2011) verwendet.

Abbildung 7.1 gibt einen Überblick über die Anteile der Missings pro Variable und die Verteilung fehlender Werte im Datensatz. Die linke Seite dieser Grafik zeigt ein Balkendiagramm mit dem Anteil fehlender Werte pro Variable, wobei auf der Ordinate die relativen Häufigkeiten der Missings abgetragen sind. Die Variablen auf der Abszisse sind absteigend nach der Höhe des Missing-Anteils geordnet. Während für die Variablen aus der ersten Erhebungswelle im Schuljahr 2004/2005 (DK3L, MK3 und DK3S) mit jeweils 59% der höchste Missing-Anteil zu verzeichnen ist, liegen hingegen für die Variablen Schulart (SART.M und SART.D) und Geschlecht (SEX) keine fehlenden Werte vor. Dieser hohe Missing-Anteil auf den fachspezifischen Leistungsvariablen DK3L, MK3 und DK3S ist vermutlich auf einen Fehler bei der Datenerfassung der Schülerstammdaten im Rahmen des Thüringer Schülerlängsschnitts im Schuljahr 2004/2005 zurückzuführen (C. Nachtigall, persönl. Mitteilung, 19.10.2010). Dadurch konnten seitens des Thüringer Landesrechenzentrums für lediglich ca. 41% der Schüler in Klassenstufe 8 eine eindeutige Zuordnung ihrer Leistungsdaten aus den Kompetenztests in Klassenstufe 3 vorgenommen werden. Auch für die Variablen MK6 und DK6 aus der zweiten Erhebungswelle ist der Anteil fehlender Werte mit 22% relativ zu den anderen Variablen hoch. Dieser Missing-Anteil lässt sich vermutlich ebenso in erster Linie auf Probleme bei der Datenzuordnung seitens des Landesrechenzentrums zurückführen.

Die rechte Seite von Abbildung 7.1 zeigt einen sog. *Aggregation plot*, welcher alle beobachteten Kombinationen bzw. Muster fehlender und beobachteter Werte sowie deren Häufigkeit darstellt. Dabei sind die beobachteten Werte blau und die fehlenden Werte rot markiert. Jede Zeile der Matrix stellt eine der insgesamt 135 beobachteten Missing-Kombinationen dar, wobei die Zeilen nach der Häufigkeit geordnet sind, mit der diese jeweils auftreten. Die oberste Zeile der Matrix zeigt ein Missing-Muster, bei dem auf den Kompetenztestvariablen aller drei Erhebungswellen (MK3, DK3L, DK3S, MK6, DK6, MK8, DK8) sowie auf den Variablen SES.M und SES.D fehlende Werte vorliegen (rote Zellen). Auf den restlichen Variablen hingegen liegen Beobachtungen

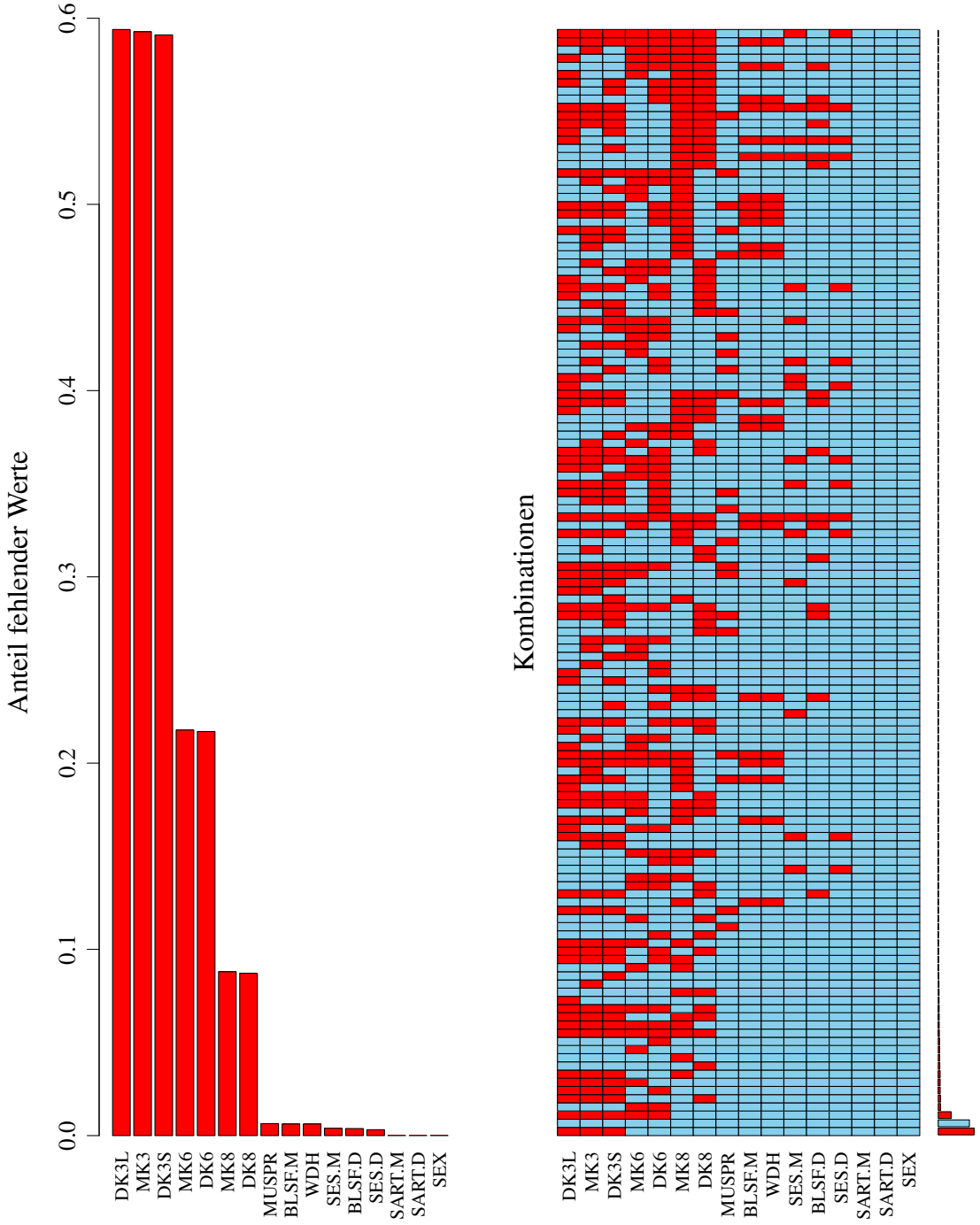

Abbildung 7.1: Missing-Struktur. Links: Balkendiagramm mit dem Anteil fehlender Werte pro Variable. Rechts: *Aggregation plot* mit allen beobachteten Kombinationen fehlender und beobachteter Werte. Beobachtete Werte sind in Blau, fehlende Werte in Rot dargestellt.

vor (blaue Zellen). Dieses ist ein seltenes Missing-Muster, welches lediglich bei einem der 12 708 Fälle auftritt. Dahingegen zeigt die letzte Zeile das häufigste Missing-Muster: Im vorliegenden Datensatz liegen bei 4 240 der 12 708 Fälle (33%) Beobachtungen auf allen Variablen außer den Variablen aus Erhebungswelle 1 (MK3, DK3L und DK3S) vor. Bei 3 687 aller Fälle (29%) liegen Beobachtungen auf allen Variablen bzw. keine fehlenden Werte vor. Ein fallweiser Ausschluss aller Fälle mit Missings auf mindestens einer der Variablen würde also zu einer Reduktion um 71% der Fälle des Datensatzes führen. Wie bereits in Abschnitt 6.3.2 erläutert, führt dieses Vorgehen – neben einem Power-Verlust – u. U. zu verzerrten Parameterschätzungen, wenn die Annahme *vollständig zufällig* fehlender Werte (MCAR) verletzt ist.

Die Frage, die sich nun stellt, lautet somit: Gibt es Hinweise, dass die Wahrscheinlichkeit des Auftretens fehlender Werte auf einer Variable von anderen Variablen abhängt? Sobald dies für mindestens eine der erhobenen Variablen zutrifft, kann nicht mehr von MCAR für den Datensatz ausgegangen werden. Um diese Annahme zu prüfen, betrachten wir nachfolgend zunächst die grafische Diagnose in Abbildung 7.2. Hier werden parallele Boxplots für die Verteilung der beobachteten Mathematikleistungsscores in Klassenstufe 8 (MK8) jeweils in Abhängigkeit von der Missing-Struktur auf allen anderen Variablen dargestellt. Auf der linken Seite von Abbildung 7.2 ist die Verteilung der Variable MK8 als Boxplot (weiß) abgebildet. Diese Verteilung wird (rechts daneben) insgesamt zehn Mal in jeweils zwei Gruppen betrachtet. Die Gruppen entsprechen dabei den Fällen mit beobachteten (blau) bzw. mit fehlenden (rot) Werten auf allen anderen Variablen des Datensatzes (DK8, MK6, DK6, MK3, DK3L, DK3S, BLSF.D, SES.M, SES.D und MUSPR). Die Prüfung der MCAR-Annahme folgt dem Falsfikationsprinzip: Aus der Annahme *vollständig zufällig* fehlender Werte (MCAR) folgt, dass es keinerlei Unterschiede in der Verteilung der Variable MK8 zwischen Schülern mit Missings und Schülern ohne Missings auf allen anderen Variablen gibt[4]. Finden sich Verteilungsunterschiede zwischen den beiden Gruppen auf *mindestens* einer anderen Variablen des Datensatzes, so ist diese MCAR-Annahme nicht mehr haltbar und muss verworfen werden. Die Variablen Schulart (SART.M und SART.D) und Geschlecht (SEX) sind nicht in der Grafik enthalten, da für diese keine fehlenden Werte vorlie-

[4]Dies muss nicht nur für die Variable MK8 gelten, sondern gleichfalls für alle anderen Variablen, wenn die MCAR-Annahme erfüllt ist. Findet man jedoch Unterschiede bezüglich mindestens einer der Variablen, so ist die Hypothese vollständig zufällig fehlender Werte falsifiziert. Zum Zwecke der Übersichtlichkeit beschränke ich mich nachfolgend auf die Darstellung des Vergleichs der Verteilungen beider Gruppen bezüglich der Variable MK8.

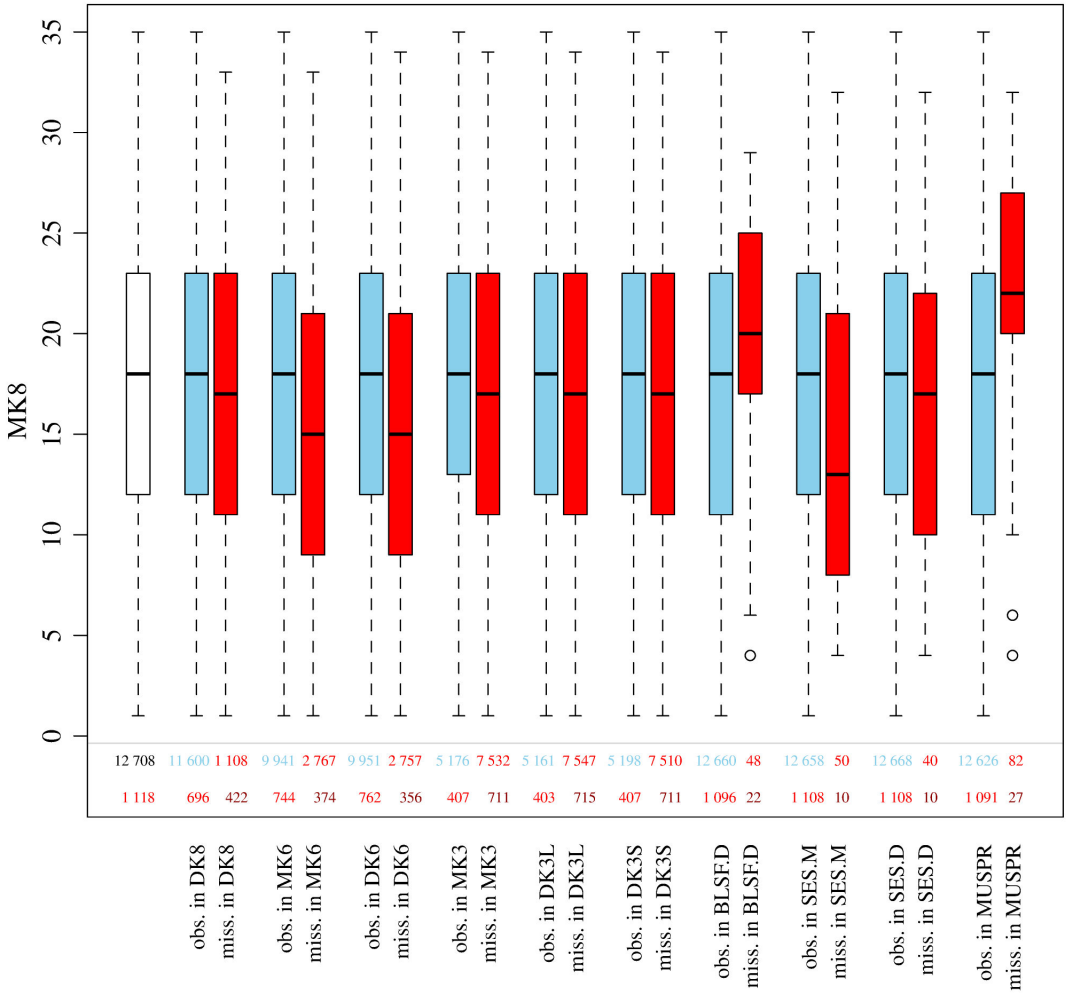

Abbildung 7.2: Der weiße Boxplot (links) zeigt die Verteilung der beobachteten Mathematikleistungsscores in Klassenstufe 8 (MK8). Rechts daneben: Parallele Boxplots für die Verteilung von MK8 in Abhängigkeit von der Missing-Struktur bezüglich der Variablen DK8, MK6, DK6, MK3, DK3L, DK3S, BLSF.D, SES.M, SES.D und MUSPR. Die Verteilung von MK8 wird hier in je zwei Gruppen dargestellt; getrennt nach dem Fehlen (rot = *missing*) und Nicht-Fehlen (blau = *observed*) auf den anderen Variablen des Datensatzes. Unterhalb der Boxplots sind die absoluten Häufigkeiten der beobachteten bzw. fehlenden Werte abgetragen.

gen. Des Weiteren sind die Variablen BLSF.M und WDH nicht aufgeführt, da aufgrund einer zu geringen Missing-Anzahl[5] keine Boxplots für diese Gruppe dargestellt werden können. Der erste blaue Boxplot stellt somit die Verteilung von MK8 für alle Fälle des Datensatzes dar, für die gleichfalls eine Beobachtung auf der Variable der DK8 vorliegt. Entsprechend zeigt der erste rote Boxplot die Verteilung von MK8 für die Fälle des Datensatzes, für welche die Werte der Variable DK8 fehlen. Beide Boxplots deuten auf symmetrische Verteilungen der Variable MK8 hin, jedoch werden Unterschiede bezüglich des Medians zwischen den beiden Gruppen sichtbar. Dieser Unterschied bezüglich des Medians von MK8 ist noch deutlicher zwischen den beiden Gruppen, die sich durch das Vorhandensein bzw. das Fehlen der Werte auf der Variable MK6 auszeichnen. Der Median der Gruppe, in der Beobachtungen für die Variable MK6 vorliegen, ist dabei größer als der in der Gruppe mit Missings auf dieser Variable. Dieser Befund zeigt sich gleichfalls für die restlichen Variablen, wobei lediglich bezüglich der Variablen Förderbedarf in Deutsch (BLSF.D) und Muttersprache (MUSPR) ein geringerer Medianwert in der Gruppe ohne Missings (im Vergleich zur Gruppe mit Missings) auf der jeweiligen Variable beobachtet wird. Im Mittel zeigen Schüler mit fehlenden Werten auf anderen Variablen somit eine geringere Mathematikleistung. Ein ähnliches Muster zeigt sich für die Leistungsscores im Fach Deutsch (vgl. Anhang D). Die Ergebnisse der grafischen Analyse indizieren somit einen leistungsabhängigen, d. h. systematischen Dropout.

Neben der grafischen Diagnose schlagen Peugh und Enders (2004) vor, für jede Variable Response-Indikatoren zu bilden (vgl. Gleichung 6.11) und die Mittelwertsunterschiede zwischen der Gruppe mit beobachteten und der Gruppe mit fehlenden Werten mittels t-Tests zu betrachten. Finden sich Mittelwertsunterschiede für mindestens eine der Variablen des Datensatzes, muss die Annahme vollständig zufällig fehlender Werte (MCAR) verworfen werden. Neben den Signifikanztests sollten dabei stets auch Effektstärkemaße angegeben werden, um die praktische Bedeutsamkeit dieser Unterschiede beurteilen zu können (Thompson, 1999). Tabelle 7.2 enthält Mittelwerte, t-Werte und Effektstärken der Mittelwertsvergleiche bezüglich der Mathematikleistung in Klassenstufe 8 (MK8). Hierbei wurden die Response-Indikatoren der zehn Variablen betrachtet, die auch der grafischen Diagnose zugrunde lagen (vgl. Abbildung 7.2). Sieben der zehn Mittelwertsunterschiede werden auf einem Signifikanzniveau von .05 signifikant. So ist bspw. der Mittelwert von MK8 der Schüler, die einen beobachteten Wert auf

[5] Hier ist nicht die Gesamtanzahl fehlender Werte auf diesen Variablen entscheidend, sondern die Anzahl der Missings auf diesen Variablen für die Fälle mit vollständigen Werten auf MK8.

Tabelle 7.2: Mittelwerte, t-Test und Effektstärken mit der Mathematikleistung in Klassenstufe 8 (MK8) als abhängige Variable und den Indikatorvariablen für fehlende Werte (Response-Indikatoren) als unabhängige Variablen

Response-Indikator[a]	M_{MK8} für $R_{[.]}$=1	M_{MK8} für $R_{[.]}$=0	t-Wert	(p-Wert)	d
R_{DK8}	17.23	16.88	1.22	(.222)	0.05
R_{MK6}	17.76	15.10	16.25	(.000)	0.37
R_{DK6}	17.72	15.24	15.02	(.000)	0.35
R_{MK3}	17.84	16.77	7.89	(.000)	0.15
R_{DK3L}	17.82	16.79	7.67	(.000)	0.14
R_{DK3S}	17.82	16.78	7.75	(.000)	0.14
$R_{BLSF.D}$	17.21	19.50	-1.78	(.087)	-0.32
$R_{SES.M}$	17.22	14.45	2.29	(.028)	0.38
$R_{SES.D}$	17.21	16.00	0.85	(.402)	0.17
R_{MUSPR}	17.19	21.73	-5.16	(.000)	-0.63

Anmerkungen. [a] Indikatorvariable $R_{[.]}$, die mit dem Wert 1 anzeigt, dass für die entsprechende Variable [.] Beobachtungen vorliegen. Der Wert 0 zeigt an, dass keine Beobachtungen (Missings) für diese Variable [.] vorliegen (vgl. Gleichung 6.11).
[a] Als Effektstärkemaß wird Cohen's d (Cohen, 1988) verwendet.

MK6 haben, signifikant höher als der durchschnittliche MK8-Score aller Schüler, die auf MK6 einen fehlenden Wert aufweisen (t = 16.25, p = .000). Die Effektstärke liegt bei d = 0.37. Dies widerspricht der Annahme, dass die Wahrscheinlichkeit fehlender Werte einer Variablen unabhängig ist von anderen Variablen im Datensatz. Die Beträge der Effektstärken für sämtliche der zehn Vergleiche reichen von 0.05 bis 0.63 (mittlerer Effekt nach Cohen, 1988). Die durchschnittliche Effektstärke beträgt d = 0.27. Diese Befunde entsprechen den Ergebnissen der grafischen Diagnose, die ebenfalls einen leistungsabhängigen, d. h. systematischen Dropout indizieren.

Da zwischen den Variablen MK8 und DK8 mit r = .68 eine starke Korrelation (Cohen, 1988) zu finden ist, zeigen sich ähnliche Ergebnisse auch für die zweite abhängige Variable DK8. Die entsprechenden Analysen für die Variable DK8 finden sich in Anhang D. Somit kann nicht von der stochastischen Unabhängkeit zwischen der Wahrscheinlichkeit fehlender Werte einer Variablen und der anderen Variablen im Datensatz ausgegangen werden. Die Annahme, dass die fehlenden Werte MCAR (d. h. *vollständig*

zufällig) sind, wird somit verworfen.

7.2 Multiple Imputation fehlender Werte

Da die fehlenden Werte nicht vollständig zufällig (MCAR) sind, führen Standardverfahren zur Behandlung fehlender Werte wie *listwise* oder *pairwise deletion* zu potenziell verzerrten Parameterschätzungen. Zum aktuellen Stand der Forschung wird die multiple Imputation (MI) als das geeignetste Verfahren zum Umgang mit fehlenden Werten erachtet (vgl. z. B. Graham, 2009; Lüdtke et al., 2007). Zur multiplen Imputation der fehlenden Werte im vorliegenden Datensatz wurde das Verfahren *Multiple Imputation by Chained Equations* (MICE; van Buuren, 2007; van Buuren & Oudshoorn, 1999; van Buuren, Brand, Groothuis-Oudshoorn & Rubin, 2006) verwendet.

7.2.1 Multiple Imputation mit MICE

Dieses Verfahren weist mehrere Vorteile[6], die es für die vorliegende Anwendung besonders prädestinieren: (a) Erstens können Variablen mit unterschiedlichen Skalenniveaus in das Imputationsmodell aufgenommen werden, da pro Variable ein anderes Imputationsmodell spezifiziert werden kann (z. B. eine logistische Regression bei einem dichotomen Regressand etc.). So können nicht nur metrische, sondern auch kategoriale (nominale) und geordnet kategoriale (ordinale) Variablen in das Imputationsmodell aufgenommen werden. Im Gegensatz zu MICE gehen andere Verfahren der multiplen Imputation zudem von einer multivariaten Normalverteilung der im Impuationsmodell verwendeten Variablen aus. Dies wiederum erfordert metrische Variablen. MICE bietet sich daher insbesondere an, wenn – wie im vorliegenden Datensatz – viele ordinale oder nominale Variablen mit fehlenden Werten vorliegen. (b) Zweitens dürfen bei der Verwendung von MICE auch bei den für die Imputation verwendeten Prädiktorvariablen im Imputationsmodell fehlende Werte vorkommen. Und schließlich (c) drittens können im Imputationsmodell zusätzlich zu den im Analysemodell verwendeten Vari-

[6] Während der erste Vorteil (a) ein Spezifikum des Verfahrens MICE ist, sind die Eigenschaften (b) und (c) gleichfalls Charakteristika von multiplen Imputationsverfahren, die auf der Annahme einer multivariaten Normalverteilung der Variablen im Imputationsmodell basieren (z. B. die MI-Implementation in der Software NORM von Schafer, 1999). Dennoch ist MICE im Rahmen der vorliegenden Anwendung besonders geeignet, da es sämtliche der drei genannten Vorteile in sich vereint.

ablen auch Hilfsvariablen (sog. *auxiliary variables*) aufgenommen werden, die einen potenziellen Bias bei der Imputation verringern können. Hilfsvariablen sind solche Variablen, die mit den Variablen im Analysemodell korrelieren oder mit dem Ausfallprozess (*missing data mechanism*) zusammenhängen (vgl. Lüdtke et al., 2007). In einer Simulationsstudie konnten Collins, Schafer und Kam (2001) zeigen, dass die Aufnahme von Hilfsvariablen, die mit dem Ausfallprozess in Zusammenhang stehen, zu einer verbesserten Parameterschätzung führt.

Hinsichtlich der Auswahl der Variablen für das Imputationsmodell geben van Buuren und Groothuis-Oudshoorn (2011b) die folgende Empfehlung: „ ... data sets often contain several hundreds of variables, all of which can potentially be used to generate imputations. It is not feasible (because of multicollinearity and computational problems) to include all these variables. ... For imputation purposes, it is expedient to select a suitable subset of data that contains no more than 15 to 25 variables“ (S. 23). Dabei sollten zunächst sämtliche Variablen in das Imputationsmodell aufgenommen werden, die auch in den späteren Analysemodellen verwendet werden. Weiterhin sollten solche (Hilfs-)Variablen aufgenommen werden, „ ... that explain a considerable amount of variance“ (van Buuren & Groothuis-Oudshoorn, 2011b, S. 23). Die Hilfsvariablen der vorliegenden Datenanalyse wurden mittels einer *Random Forest Analyse* (Breiman, 2001; Liaw & Wiener, 2002) ausgewählt. Dabei handelt es sich um ein nonparametrisches Verfahren, anhand dessen die Bedeutsamkeit verschiedener (zunächst potenzieller) Prädiktorvariablen für die Vorhersage einer abhängigen Variablen quantifiziert wird. Als Maß für die Bedeutsamkeit einer Prädiktorvariable wird die Veränderung des mittleren Fehlerquadrats (*increase in mean square error*; IncMSE) pro Variable betrachtet. Dieser Wert gibt für jede Variable an, um wie viel das mittlere Fehlerquadrat prozentual ansteigt, wenn diese Variable nicht im Vorhersagemodell enthalten ist. Der Vorteil dieses Verfahrens liegt darin, dass hierbei auch sämtliche Interaktion der Variablen, die potenziell als Prädiktoren verwendet werden, berücksichtigt werden. Aus dem vorliegenden Datensatz wurden mittels der Random Forest Analyse neun Variablen als zusätzliche Hilfsvariablen[7] für das Imputationsmodell identifiziert. Hier wurden die Variablen ausgewählt, die im Rahmen der Random Forest Analyse die größte Veränderung

[7]Folgende Hilfsvariablen wurden in das Imputationsmodell aufgenommen: Die Halbjahresnoten in Klassenstufe 3 der Fächer Mathematik und Deutsch (MK3.NOTE, DK3.NOTE), die Halbjahresnoten der Fächer Mathematik, Deutsch und Englisch der Klassenstufen 6 und 8 (MK6.NOTE, DK6.NOTE, EK6.NOTE, MK8.NOTE, DK8.NOTE, EK8.NOTE) sowie die Kompetenztest-Leistungsscores aus Klassenstufe 8 im Fach Englisch (EK8).

des mittleren Fehlerquadrats (IncMSE) aufwiesen.

Entsprechend der Empfehlung von van Buuren und Groothuis-Oudshoorn (2011b) wurden insgesamt 25 Variablen in das Imputationsmodell aufgenommen. Von diesen 25 Variablen wurden 16 Variablen (vgl. Kapitel 6; Tabelle 6.3) in den verschiedenen Analysemodellen verwendet. Imputiert wurden $m = 5$ vollständige Datensätze [8] mittels des R-Packets `MICE` (van Buuren & Groothuis-Oudshoorn, 2011a). Auf jeden der fünf Datensätze wurden sämtliche der in Kapitel 6 dargestellten Modelle (vgl. Tabelle 6.4) angewendet. Die Ergebnisse aus den fünf imputierten Datensätzen wurden schließlich nach dem von Rubin (1987) beschriebenen Verfahren zu einer Gesamtschätzung kombiniert (vgl. Anhang E).

7.2.2 Diagnostik der multiplen Imputation

Bevor wir uns den Ergebnissen des Modellvergleichs zuwenden, soll an dieser Stelle noch ein Blick auf die Plausibilität der imputierten Datensätze bzw. des Imputationsmodells geworfen werden. Bei der multiplen Imputation wird die Annahme gemacht, dass die fehlenden Werte MAR (*missing at random*) sind (vgl. Kapitel 6, Abschnitt 6.3.2). Diese Hypothese kann empirisch nicht getestet werden. Jedoch kann man prüfen, ob die auf Basis des Imputationsmodells imputierten Werte plausibel sind. Hierfür eignen sich wiederum grafische Diagnostiken.

Eine übliche Methode, die Plausibilität der Imputationen zu prüfen, besteht darin, die beobachteten Daten (mit fehlenden Werten) und die augmentierten Daten (nach der Imputation) anhand verschiedener Statistiken wie bspw. nonparametrischer Dichteschätzungen zu vergleichen (vgl. Abayomi, Gelman & Levy, 2008; Raghunathan & Bondarenko, 2007; van Buuren & Groothuis-Oudshoorn, 2011b). Abweichungen zwischen den Dichteschätzungen sind unter der MAR-Annahme nicht unwahrscheinlich und indizieren somit nicht zwingend eine Problem mit dem Imputationsmodell. Jedoch zeigen Abayomi et al. (2008) an einem hypothetischen Beispiel, dass „ . . . dramatic differences between the imputed and observed data can suggest a *potential* problem and,

[8] Hinsichtlich der Anzahl der Imputationen findet sich in der Literatur zur multiplen Imputation die allgemeine Empfehlung, dass drei bis zehn imputierte Datensätze ausreichend sind (vgl. Peugh & Enders, 2004; Rubin, 1987; Schafer, 1997). In empirischen Anwendungen findet man daher häufig $m = 5$ imputierte Datensätze (z. B. Lehmann & Lenkeit, 2008; Maaz, Baumert, Gresch & McElvany, 2010). Weitere Forschungsergebnisse zeigen, dass durch eine größere Anzahl von Imputationen die statistische Power erhöht werden kann (vgl. Baraldi & Enders, 2010; Graham, Olchowski & Gilreath, 2007), die jedoch im Rahmen der vorliegenden Arbeit nicht im Fokus der Betrachtung liegt.

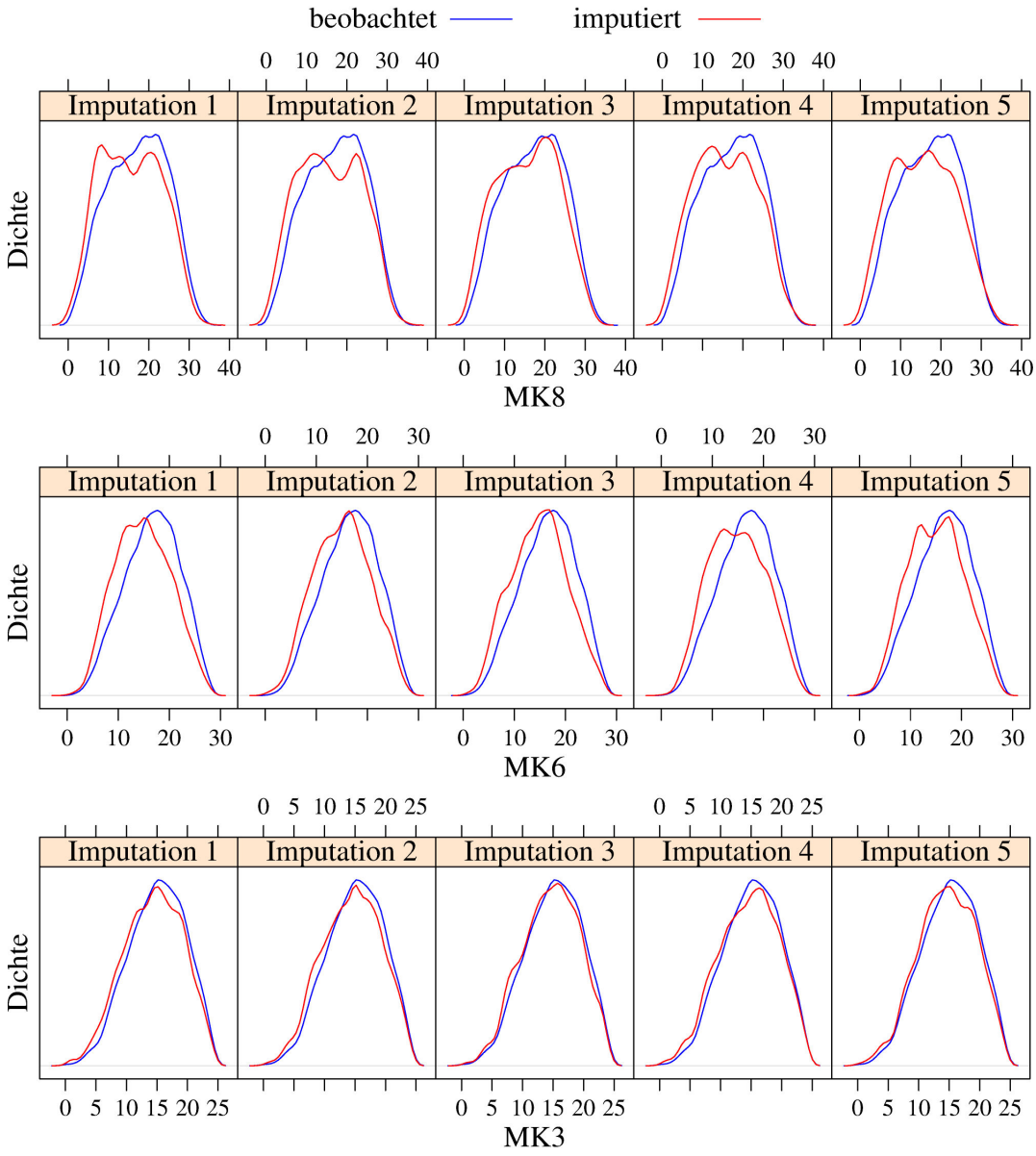

Abbildung 7.3: Nonparametrische Dichteschätzungen aufgrund der beobachteten und imputierten Werte der Mathematikleistung in Klassenstufe 3, 6 und 8 (MK3, MK6 und MK8).

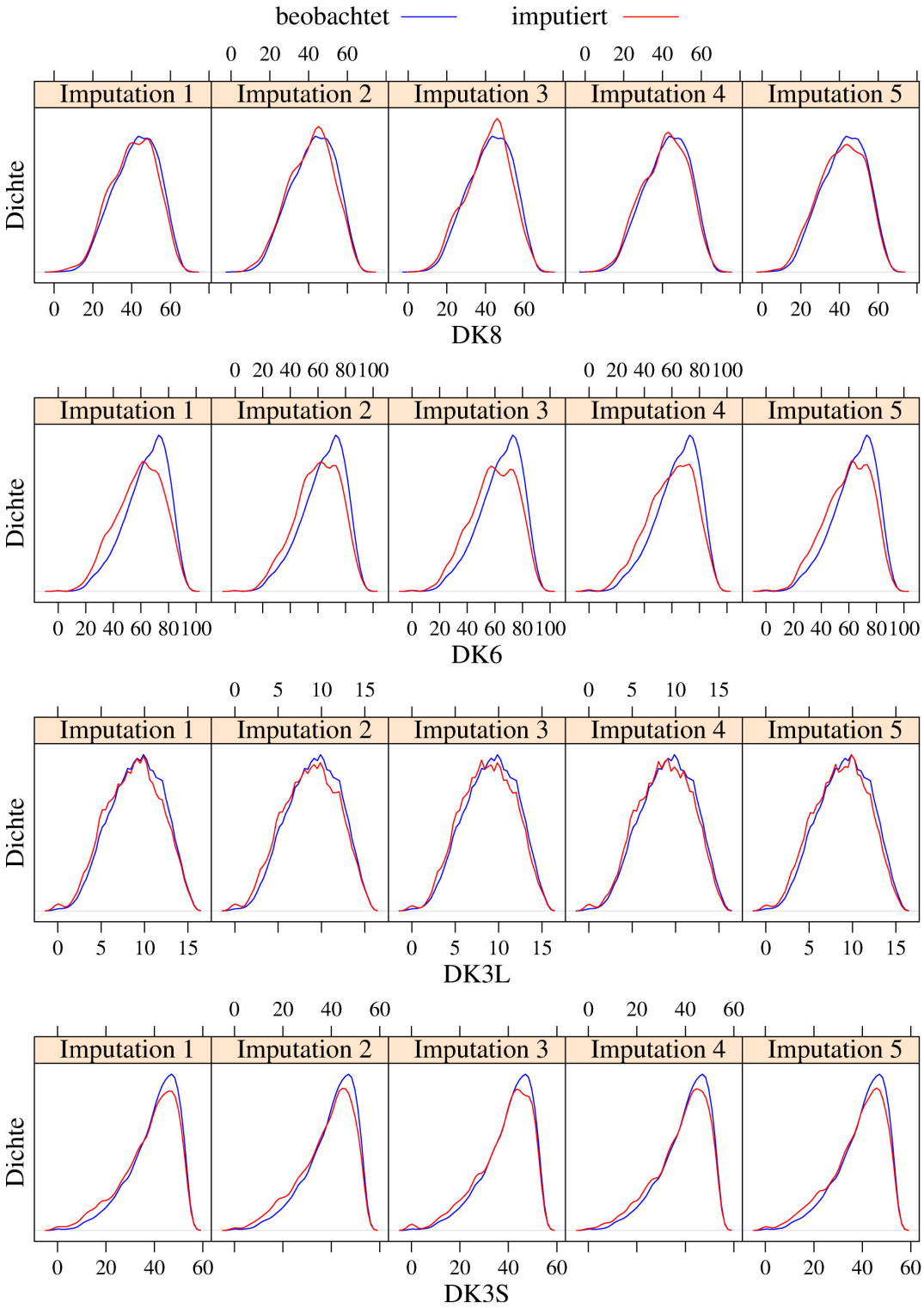

Abbildung 7.4: Nonparametrische Dichteschätzungen aufgrund der beobachteten und imputierten Werte der Deutschleistung in Klassenstufe 3, 6 und 8 (DK3L, DK3S, DK6 und DK8).

in a context with many imputed variables, it is helpful to have some screening device to identify these potential problems“ (S. 280). Dieses *screening device* soll daher auch auf den vorliegenden Datensatz angewendet werden: In den Abbildungen 7.3 und 7.4 sind die nonparametrischen Dichteschätzungen aufgrund der beobachteten und imputierten Werte für jede der fünf Imputationen sowohl für die Mathematik- als auch für die Deutschleistungen dargestellt. Bei allen Verteilungen sind nur marginale Unterschiede sichtbar. Die Unterschiede werden zudem umso geringer, je größer der Missinganteil einer Variablen ist (z. B. bei MK3 im Vergleich zu MK6 und MK8). Dieses Ergebnismuster zeigt sich konsistent für alle im Imputationsmodell aufgenommenen Variablen. Es finden sich somit keine Hinweise für Probleme bezüglich des verwendeten Imputationsmodells.

7.3 Modellvergleich

Nach der erfolgreichen Imputation der fehlenden Werte wurden für jeden der $m = 5$ imputierten Datensätze die in Kapitel 6 beschriebenen Modelle (vgl. Tabelle 6.4) berechnet; sowohl mit der Mathematikleistung in Klassenstufe 8 (MK8) als auch mit der Deutschleistung in Klassenstufe 8 (DK8) als abhängige Variable. Anschließend erfolgte die Aggregation der Ergebnisse aus den einander entsprechenden fünf imputierten Datensätzen zu einer Gesamtschätzung nach dem von Rubin (1987) beschriebenen Verfahren. Die Ergebnisse aus den 14 Modellen werden im vorliegenden Abschnitt anhand folgender Kriterien bzw. grafischer Darstellungsformen im Hinblick auf die in Kapitel 5 postulierten Hypothesen vergleichend betrachtet:

(1) Caterpillar-Plots der adjustierten klassenspezifischen Effektschätzungen,

(2) Determinationskoeffizient $R^2_{Y|\boldsymbol{Z}}$,

(3) Korrelationen der adjustierten klassenspezifischen Effektschätzungen beim paarweisen Modellvergleich,

(4) Change-Plots der Veränderungen der Effektschätzungen beim paarweisen Modellvergleich und

(5) Transitionsmatrizen, die die Veränderungen des Quintil-Rankings der Effektschätzungen beim paarweisen Modellvergleich abbilden.

Zum Zwecke der Übersichtlichkeit werden entsprechend der Modellselektion (Parametrisierung) verschiedene Farben für die grafische Darstellung der Ergebnisse verwendet: Die Ergebnisse der Modelle 1 bis 7 (saturierte bzw. bedingt lineare Parametrisierung) werden in der Farbe Cyan dargestellt, während die Ergebnisse der Modelle 8 bis 14 (lineare Parametrisierung) in der Farbe Rot dargestellt werden.

7.3.1 Caterpillar-Plots

Nachfolgend werden die adjustierten klassenspezifischen Effektschätzungen aus den in Kapitel 6 beschriebenen 14 Modellen (vgl. Tabelle 6.4) sowohl für das Fach Mathematik als auch für das Fach Deutsch dargestellt. Zur grafischen Darstellung werden zunächst sog. *Caterpillar-Plots* verwendet. Diese zeigen die Punktschätzungen der adjustierten klassenspezifischen Effekte $\bar{\delta}_{adj;x}$ sowie die zugehörigen Standardfehler[9] $SE(\bar{\delta}_{adj;x})$. Zudem veranschaulichen Caterpillar-Plots die Rangordnung der Klassen gemäß der Größe bzw. dem Ausmaß des klassenspezifischen Effekts. Auf der Ordinatenachse sind die adjustierten klassenspezifischen Effekte $\bar{\delta}_{adj;x}$ abgetragen. Die klassenspezifischen Effekte $\bar{\delta}_{adj;x}$ schwanken um den Wert null, da dies der Erwartungswert der klassenspezifischen Effekte ist (vgl. Kapitel 3; Gleichung 3.32). Dieser Erwartungswert ist in den Abbildungen 7.5 bis 7.10 jeweils mit einer gestrichelten grauen Linie gekennzeichnet, die parallel zur Abszisse verläuft. Die Abszisse repräsentiert die einzelnen Klassen, die hinsichtlich der Größe der adjustierten klassenspezifischen Effekte in eine Rangreihe gebracht wurden. Auf diese Weise gibt die Position einer Klasse auf der Abszisse gleichfalls deren Rang wieder. Zudem erhält der Caterpillar-Plot durch das Ranking der Punktschätzer bei gleichzeitiger Angabe der Standardfehler sein raupenähnliches Aussehen, welchem er seinen Namen verdankt (*Raupe*; engl.: caterpillar).

In jeder der Abbildungen 7.5 bis 7.10 wird zusätzlich der jeweilige prozentuale Anteil der Klassen angegeben, der bei einem Signifikanzniveau von $\alpha = .05$ signifikant von null abweicht. Zudem werden in den Tabellen 7.3 und 7.4 die Varianzen $s^2(\bar{\delta}_{adj;x})$ der adjustierten Effektschätzungen im Fach Mathematik sowie Deutsch wiedergegeben.

[9] Die Berechnung der Standardfehler der adjustierten klassenspezifischen Effektschätzungen $SE(\bar{\delta}_{adj;x})$ ist im Anhang E formal dargestellt.

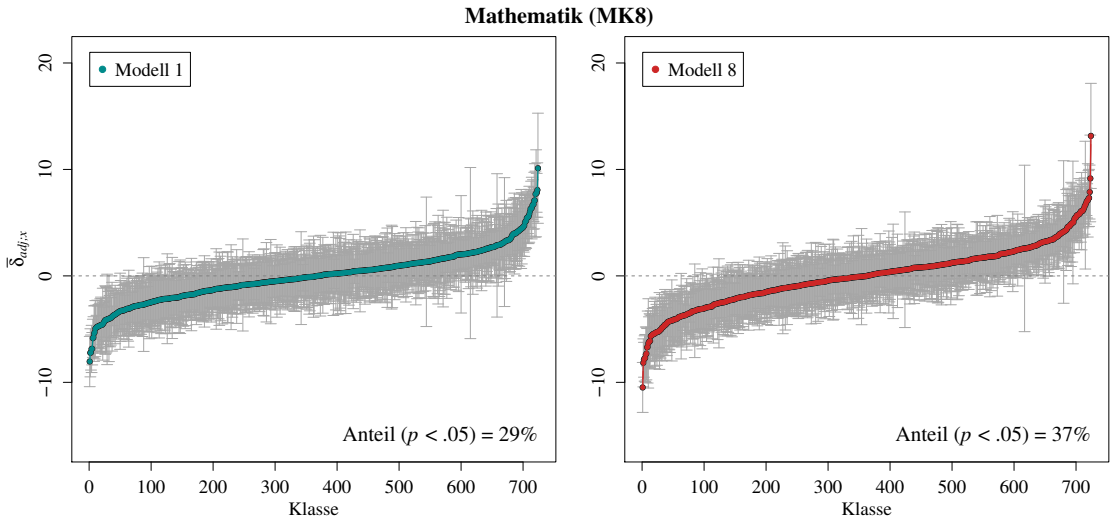

Abbildung 7.5: Caterpillar-Plots der CAM im Fach Mathematik (MK8). Links: Saturierte Parametrisierung (Modell 1). Rechts: Lineare Parametrisierung ohne Interaktionen (Modell 8).

Caterpillar-Plots im Fach Mathematik

Die Caterpillar-Plots der adjustierten Effektschätzungen im Fach Mathematik sind in den Abbildungen 7.5 bis 7.7 dargestellt. Tabelle 7.3 zeigt die Varianzen der adjustierten Effektschätzungen pro Modell in der Übersicht.

CAM. Abbildung 7.5 zeigt die Caterpillar-Plots der Contextualized Attainment Modelle (CAM) im Fach Mathematik (MK8) sowohl für die saturierte Parametrisierung (Modell 1) als auch für die lineare Parametrisierung ohne Interaktionen (Modell 8). Die Varianz der Effektschätzungen aus Modell 1 ist mit $s^2_{M1}(\bar{\delta}_{adj;x}) = 5.70$ niedriger als die Varianz der Effektschätzungen aus Modell 8 mit $s^2_{M8}(\bar{\delta}_{adj;x}) = 7.76$. Des Weiteren ist der Anteil der Klassen, deren Effektschätzung signifikant von null verschieden ist, bei Modell 1 geringer als bei Modell 8 (29% vs. 38%).

VAM. In Abbildung 7.6 werden die Ergebnisse der Value-Added Modelle (VAM) für beide Parametrisierungen dargestellt. Die Ergebnisse aus Modellen mit bedingt linearer Parametrisierung (mit Interaktionen) sind auf der linken Seite abgebildet und rechts sind die Ergebnisse der linear parametrisierten Modelle ohne Interaktionen dargestellt. Auch hier zeigt sich, dass die Varianzen der Effektschätzungen aus den komplexeren

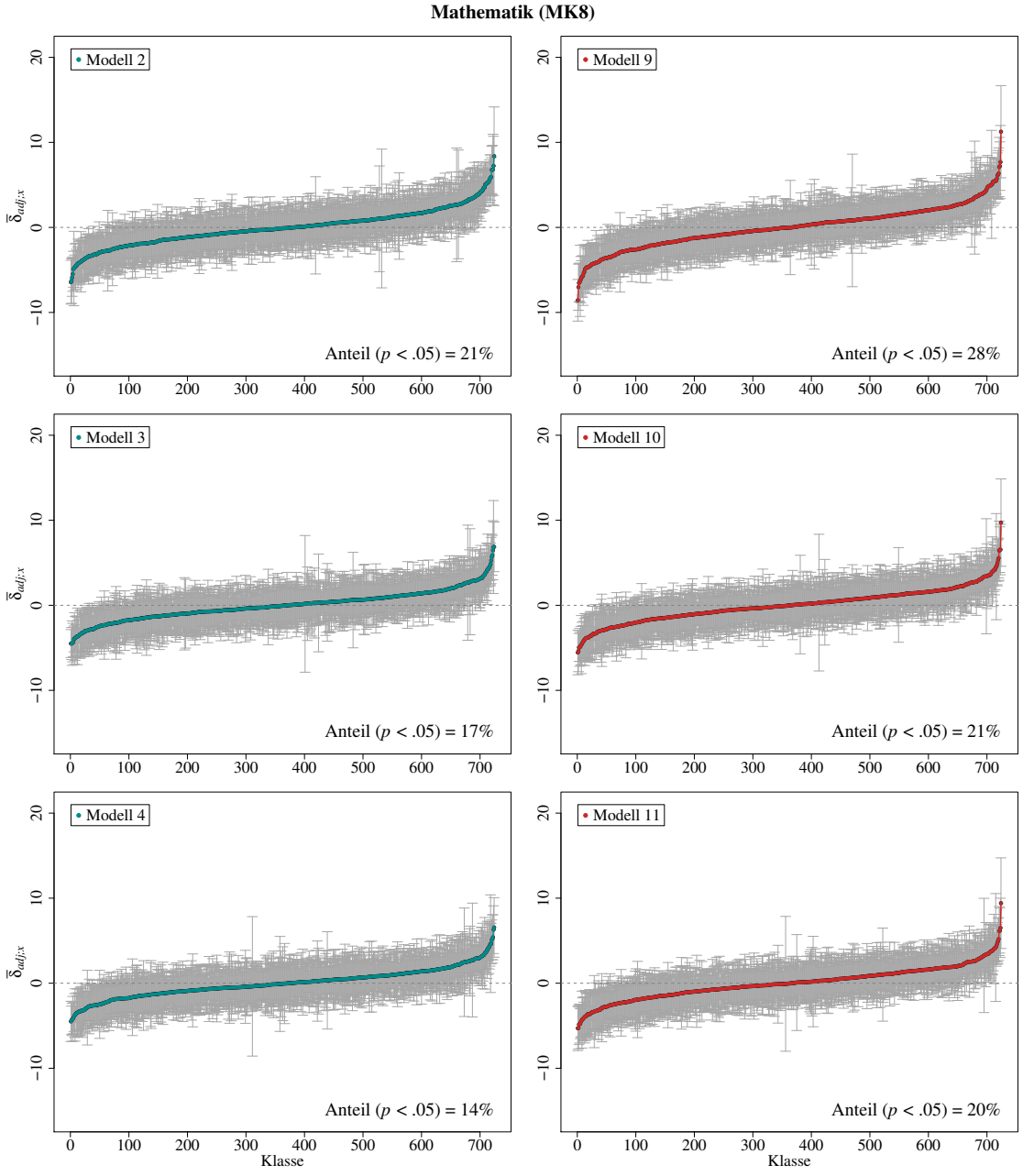

Abbildung 7.6: Caterpillar-Plots der VAM im Fach Mathematik (MK8). Links: Bedingt lineare Parametrisierung mit Interaktionen (Modelle 2, 3, 4). Rechts: Lineare Parametrisierung ohne Interaktionen (Modelle 9, 10, 11).

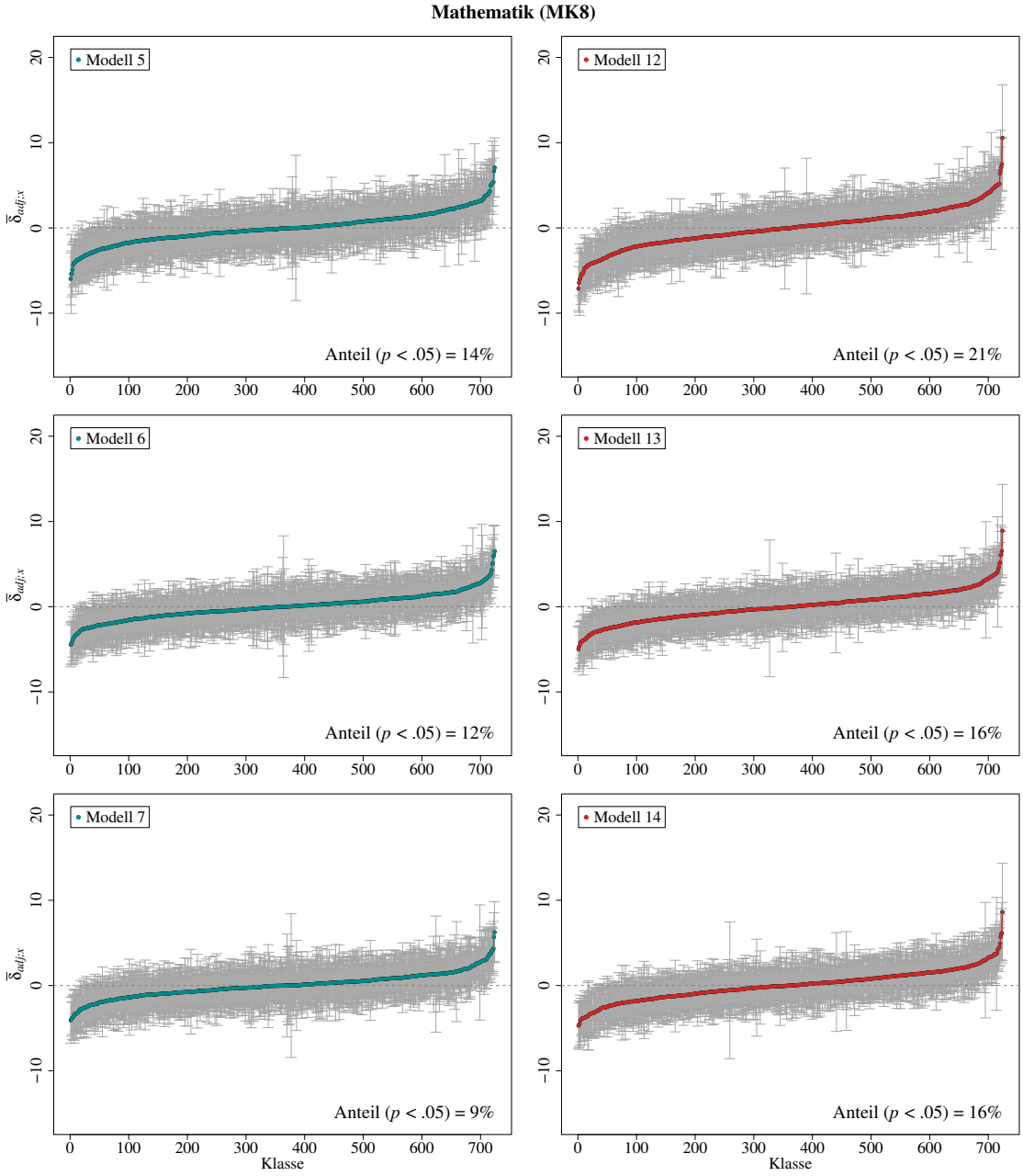

Abbildung 7.7: Caterpillar-Plots der CVA im Fach Mathematik (MK8). Links: Bedingt lineare Parametrisierung mit Interaktionen (Modelle 5, 6, 7). Rechts: Lineare Parametrisierung ohne Interaktionen (Modelle 12, 13, 14).

Tabelle 7.3: Varianz der Effektschätzungen $s^2(\bar{\delta}_{adj;x})$ pro Modell im Fach Mathematik (MK8)

Modellselektion	Kovariatenselektion						
	CAM	VAM			CVA		
bedingt lineare Parametrisierung (inkl. Interaktionen)	$M1^a$	$M2$	$M3$	$M4$	$M5$	$M6$	$M7$
	5.70	4.29	2.89	2.65	3.06	2.21	1.91
lineare Parametrisierung (ohne Interaktionen)	$M8$	$M9$	$M10$	$M11$	$M12$	$M13$	$M14$
	7.76	5.60	3.55	3.39	4.79	3.04	2.91

Anmerkungen. CAM = Contextualized Attainment Model, VAM = Value-Added Model, CVA = Contextual Value-Added Model, M = Modell.
[a] Saturiertes Zellenmittelwertemodell.

Modellen (bedingt lineare Parametrisierung; Modelle 2, 3 und 4) jeweils geringer sind, als die Varianzen aus den hinsichtlich der Kovariatenselektion entsprechenden Modellen 9, 10 und 11 mit linearer Parametrisierung: $s^2_{M2}(\bar{\delta}_{adj;x}) = 4.29$, $s^2_{M3}(\bar{\delta}_{adj;x}) = 2.89$, $s^2_{M4}(\bar{\delta}_{adj;x}) = 2.65$ vs. $s^2_{M9}(\bar{\delta}_{adj;x}) = 5.60$, $s^2_{M10}(\bar{\delta}_{adj;x}) = 3.55$, $s^2_{M11}(\bar{\delta}_{adj;x}) = 3.39$. Gleichsam ist auch hier der Anteil der Klassen, deren Effektschätzung signifikant von null verschieden ist, bei den Modellen 2, 3 bzw. 4 (21%, 17% und 14%) jeweils geringer als bei den entsprechenden Modellen 9, 10 und 11 (28%, 21% und 20%).

CVA. Die in Abbildung 7.7 dargestellten Ergebnisse der Contextual Value-Added Modelle (CVA) zeigen ein ganz ähnliches Ergebnismuster. Die Varianzen aus den Modellen 5, 6 und 7 (bedingt lineare Parametrisierungen mit Interaktionen) sind kleiner, als die Varianzen aus den entsprechenden linear-parametrisierten Modellen 12, 13 und 14 ohne Interaktionen: $s^2_{M5}(\bar{\delta}_{adj;x}) = 3.06$, $s^2_{M6}(\bar{\delta}_{adj;x}) = 2.21$, $s^2_{M7}(\bar{\delta}_{adj;x}) = 1.91$ vs. $s^2_{M12}(\bar{\delta}_{adj;x}) = 4.79$, $s^2_{M13}(\bar{\delta}_{adj;x}) = 3.04$, $s^2_{M14}(\bar{\delta}_{adj;x}) = 2.91$. Ebenso ist der Anteil der Klassen mit signifikanten Effektschätzungen bei den Modellen mit bedingt linearer Parametrisierung (14%, 12% und 9%) geringer als bei den entsprechenden linearen Modellen (21%, 16% und 16%).

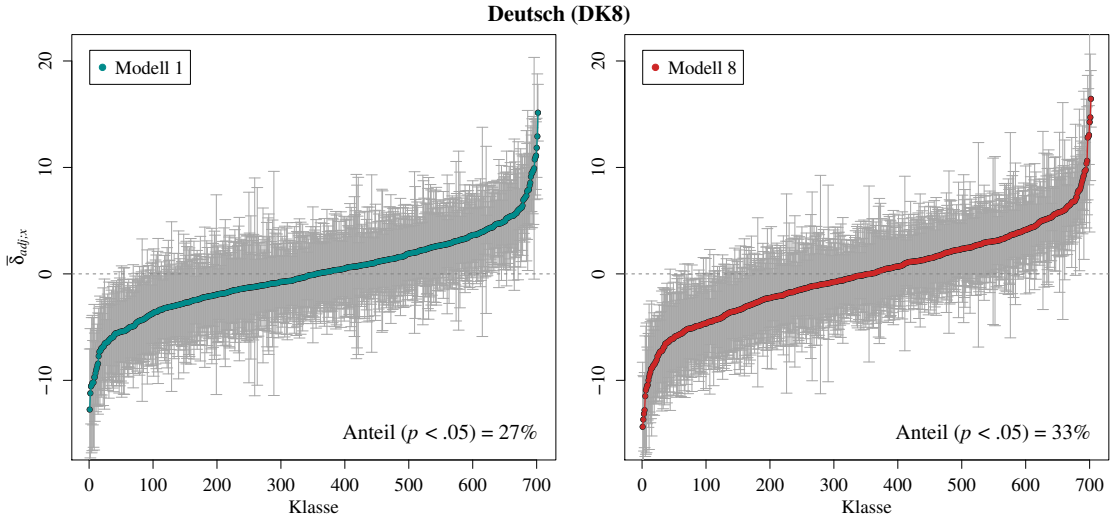

Abbildung 7.8: Caterpillar-Plots der CAM im Fach Deutsch (DK8). Links: Saturierte Parametrisierung (Modell 1). Rechts: Lineare Parametrisierung ohne Interaktionen (Modell 8).

Caterpillar-Plots im Fach Deutsch

Die Abbildungen 7.8 bis 7.10 zeigen die Caterpillar-Plots der adjustierten Effektschätzungen im Fachbereich Deutsch. Tabelle 7.4 zeigt die Varianzen der Effektschätzungen $s^2(\bar{\delta}_{adj;x})$ im Fach Deutsch pro Modell in der Übersicht.

CAM. Abbildung 7.8 zeigt die Caterpillar-Plots der Contextualized Attainment Modelle (CAM) im Fach Deutsch (DK8); wiederum sowohl für die saturierte Parametrisierung (Modell 1) als auch für die lineare Parametrisierung ohne Interaktionen (Modell 8). Die Varianz der Effektschätzungen aus Modell 1 ist mit $s^2_{M1}(\bar{\delta}_{adj;x}) = 13.84$ niedriger als die Varianz der Effektschätzungen aus Modell 8 mit $s^2_{M8}(\bar{\delta}_{adj;x}) = 18.00$. Auch im Fach Deutsch ist der Anteil der Klassen, deren Effektschätzung signifikant von null verschieden ist, bei Modell 1 geringer als bei Modell 8 (27% vs. 33%).

VAM. In Abbildung 7.9 sind die Ergebnisse der Value-Added Modelle (VAM) im Fach Deutsch für beide Parametrisierungen dargestellt. Auch hier zeigt sich, dass die Varianzen der Effektschätzungen aus den komplexeren Modellen (bedingt lineare Parametrisierung inkl. Interaktionen) Modelle 2, 3 und 4 mit $s^2_{M2}(\bar{\delta}_{adj;x}) = 9.72$, $s^2_{M3}(\bar{\delta}_{adj;x}) = 9.30$ und $s^2_{M4}(\bar{\delta}_{adj;x}) = 7.89$ jeweils geringer sind, als die Varianzen aus den bezüglich

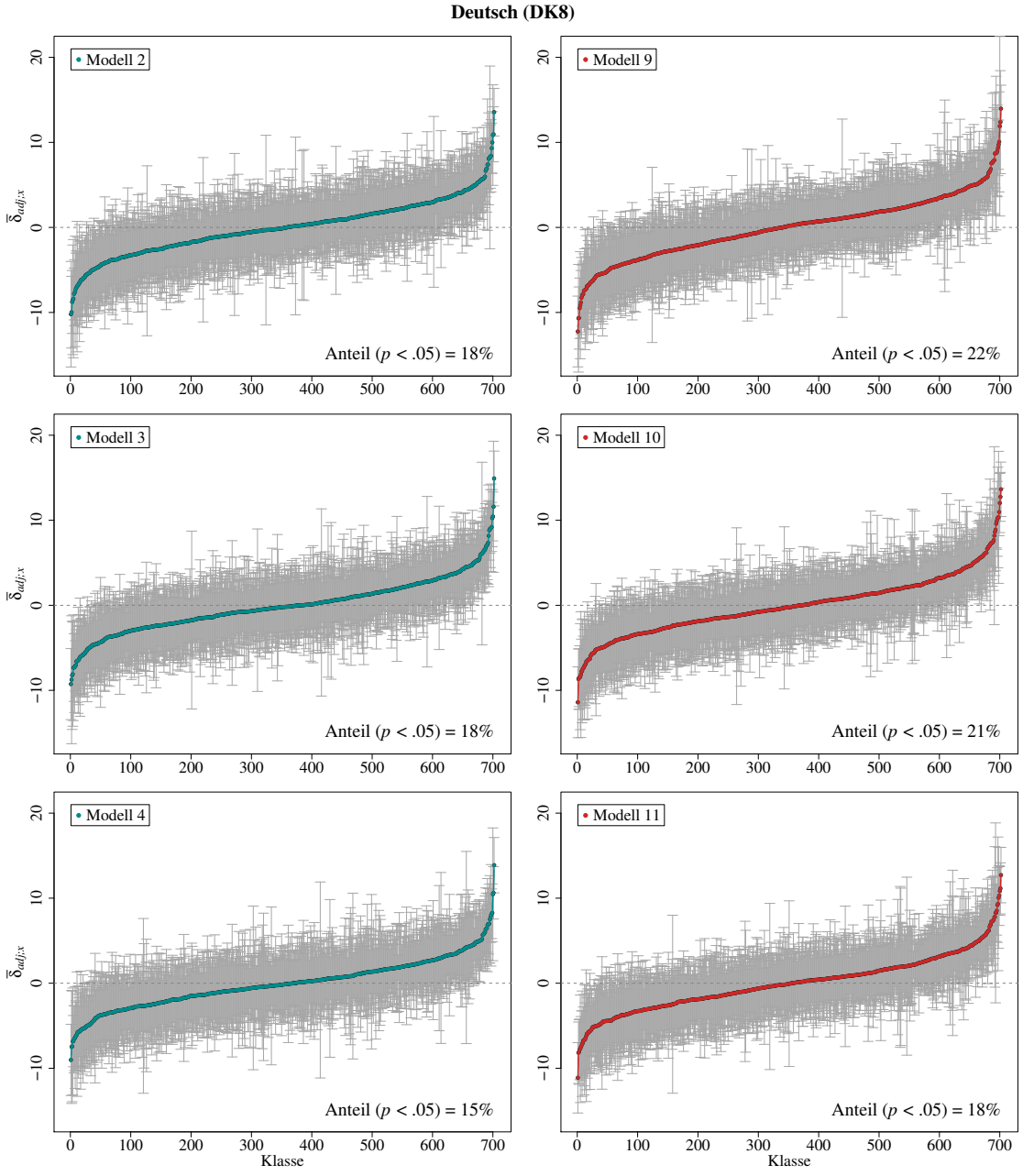

Abbildung 7.9: Caterpillar-Plots der VAM im Fach Deutsch (DK8). Links: Bedingt lineare Parametrisierung mit Interaktionen (Modelle 2, 3, 4). Rechts: Lineare Parametrisierung ohne Interaktionen (Modelle 9, 10, 11).

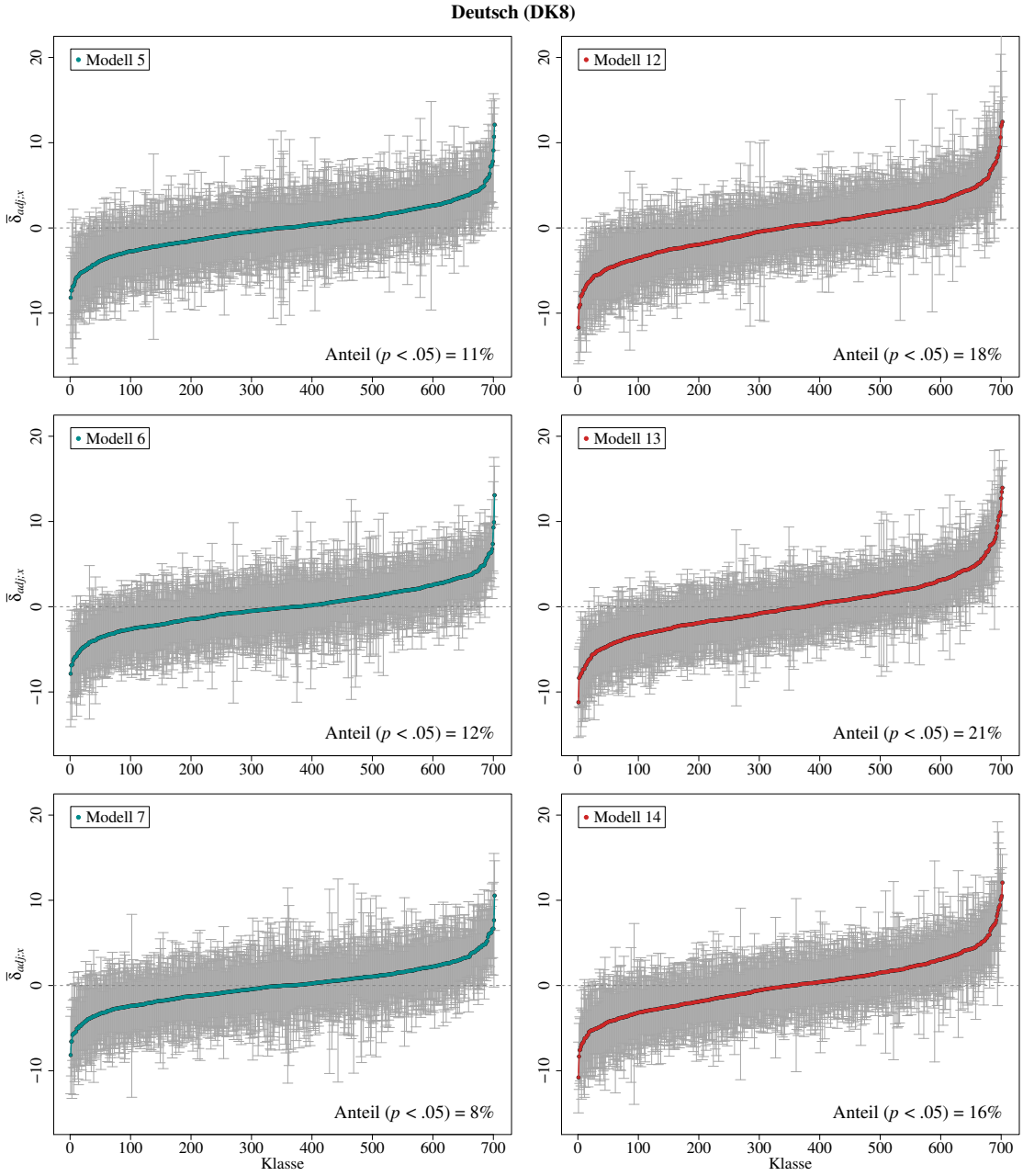

Abbildung 7.10: Caterpillar-Plots der CVA im Fach Deutsch (DK8). Links: Bedingt lineare Parametrisierung mit Interaktionen (Modelle 5, 6, 7). Rechts: Lineare Parametrisierung ohne Interaktionen (Modelle 12, 13, 14).

Tabelle 7.4: Varianz der Effektschätzungen $s^2(\bar{\delta}_{adj;x})$ pro Modell im Fach Deutsch (DK8)

Modellselektion	Kovariatenselektion						
	CAM	VAM			CVA		
bedingt lineare Parametrisierung (inkl. Interaktionen)	$M1^a$	$M2$	$M3$	$M4$	$M5$	$M6$	$M7$
	13.84	9.72	9.30	7.89	7.09	6.51	5.38
lineare Parametrisierung (ohne Interaktionen)	$M8$	$M9$	$M10$	$M11$	$M12$	$M13$	$M14$
	18.00	12.30	11.16	9.94	11.05	11.09	9.52

Anmerkungen. CAM = Contextualized Attainment Model, VAM = Value-Added Model, CVA = Contextual Value-Added Model, M = Modell.
[a] Saturiertes Zellenmittelwertemodell.

der Kovariatenselektion entsprechenden Modellen 9, 10 und 11 mit linearer Parametrisierung ohne Interaktionen. Die zuletzt genannten Varianzen sind $s^2_{M9}(\bar{\delta}_{adj;x}) = 12.30$, $s^2_{M10}(\bar{\delta}_{adj;x}) = 11.16$ und $s^2_{M11}(\bar{\delta}_{adj;x}) = 9.94$. Auch hier ist der Anteil der Klassen, deren Effektschätzung signifikant von null verschieden ist, bei den Modellen 2, 3 bzw. 4 (18%, 18% und 15%) jeweils geringer als bei den entsprechenden Modellen 9, 10 und 11 (22%, 21% und 18%).

CVA. Auch die in Abbildung 7.10 dargestellten Ergebnisse der Contextual Value-Added Modelle (CVA) zeigen ein vergleichbares Ergebnismuster. Die Varianzen aus den Modellen 5, 6 und 7 (bedingt lineare Parametrisierungen mit Interaktionen) betragen $s^2_{M5}(\bar{\delta}_{adj;x}) = 7.09$, $s^2_{M6}(\bar{\delta}_{adj;x}) = 6.51$ und $s^2_{M7}(\bar{\delta}_{adj;x}) = 5.38$. Diese sind jeweils kleiner als die Varianzen aus den entsprechenden linear-parametrisierten Modellen 12, 13 und 14 ohne Interaktionen: $s^2_{M12}(\bar{\delta}_{adj;x}) = 11.05$, $s^2_{M13}(\bar{\delta}_{adj;x}) = 11.09$ und $s^2_{M14}(\bar{\delta}_{adj;x}) = 9.52$. Gleichsam ist der Anteil der Klassen mit signifikanten Effektschätzungen bei den Modellen mit bedingt linearer Parametrisierung (11%, 12% und 8%) geringer als bei den entsprechenden linearen Modellen (18%, 21% und 16%).

Zusammenfassung: Caterpillar-Plots

Sowohl im Fach Mathematik als auch im Fach Deutsch ist die Varianz der Effektschätzungen sowie der Anteil der Klassen mit signifikant von null abweichenden Effektschätzungen umso kleiner, je differenzierter das Modell hinsichtlich der Parametrisierung *und* der Kovariatenselektion ist. Innerhalb einer Parametrisierungsform – d. h. für beide Parametrisierungsformen – gilt jeweils: Je mehr Kovariaten in das Modell aufgenommen werden, desto geringer ist die Varianz der Effektschätzungen *und* desto geringer ist der Anteil der Klassen mit signifikant vom Erwartungswert abweichenden Effektschätzungen. Weiterhin zeigt der Vergleich zwischen den beiden Parametrisierungsformen, dass die Varianz der adjustierten Effektschätzungen in den bedingt linearen Modellen (mit Interaktionen) stets geringer ist als in den hinsichtlich des Kovariatensets entsprechenden Modellen mit linearer Parametrisierung (ohne Interaktionen). Gleichzeitig ist auch der Anteil signifikant von null abweichender Effekte jeweils geringer für die hinsichtlich der Parametrisierung komplexeren Modelle.

7.3.2 Determinationskoeffizient $R^2_{Y|Z}$

Sei $E(Y \mid X)$ eine Regression des Regressanden Y auf den Regressor X, wobei X ein zunächst beliebiger Regressor ist. Der Determinationskoeffizient[10] $R^2_{Y|X}$ der Regression $E(Y \mid X)$ ist dann wie folgt definiert (Steyer, 2003, S. 89):

$$R^2_{Y|X} \equiv \frac{Var[\, E(Y \mid X)\,]}{Var(Y)}\,, \qquad \text{falls } Var(Y) > 0\,. \tag{7.1}$$

Der Wertebereich des Determinationskoeffizienten $R^2_{Y|X}$ liegt zwischen 0 und 1. Er nimmt den Wert 0 an, falls Y von X regressiv unabhängig ist, d. h. wenn gilt: $E(Y \mid X) = E(Y)$. In diesem Fall gilt: $Var[\, E(Y \mid X)\,] = Var[\, E(Y)\,] = 0$ und der Dividend in Gleichung 7.1 nimmt somit den Wert 0 an. Er nimmt den Wert 1 an, falls Y vollständig von X abhängt, denn dann gilt: $E(Y \mid X) = Y$. In diesem Fall kann die Varianz des Regressanden Y vollständig durch X erklärt werden und für die Residualvarianz gilt: $Var(\varepsilon_{Y|Z}) = 0$. Dividend und Divisor in Gleichung 7.1 nehmen folglich den gleichen Wert – nämlich $Var(Y)$ – an, so dass gilt: $R^2_{Y|X} = 1$. Somit ist der Determinationskoeffizient $R^2_{Y|X}$ interpretierbar als „... der durch X determinierte Varianzanteil von Y“ (Steyer, 2003, S. 89).

[10] Der Determinationskoeffizient $R^2_{Y|X}$ wird auch als *Bestimmheitsmaß* der Regression $E(Y \mid X)$ bezeichnet (z. B. Fahrmeier, Kneib & Lang, 2007). Beide Begriffe sind synonym.

Multpliziert man den Determinationskoeffizient mit 100, so erhält man den Prozentsatz erklärter Varianz an der Gesamtvarianz des Regressanden Y.

Im Rahmen der vorliegenden Arbeit betrachte ich jedoch keine beliebige Regression $E(Y \mid X)$, sondern die Regression $E(Y \mid \boldsymbol{Z})$ der Testwertvariablen Y auf den mehrdimensionalen Regressor $\boldsymbol{Z}$, der die verschiedenen Kovariaten umfasst. Nachfolgend werden die Determinationskoeffizienten $R^2_{Y|\mathbf{Z}}$ der 14 Modelle aufgelistet und miteinander verglichen; zunächst für das Fach Mathematik und daran anschließend für das Fach Deutsch. Ein Modellvergleich hinsichtlich des Determinationskoeffizienten $R^2_{Y|\mathbf{Z}}$ ist jedoch nur für sog. *genestete Modelle* bedeutsam. Bollen (1989) definiert genestete Modelle wie folgt: „In general, any model which requires that some function of its free parameters equals another free paramater or equals a constant [z. B. dem Wert 0] is nested in the identical model that has no such restriction“ (Bollen, 1989, S. 291). Lässt sich ein (allgemeineres) Modell durch die Restriktion von Parametern in ein anderes (restriktiveres) Modell überführen, so sind diese Modelle ineinander genestet. So ist bspw. Modell 1 ein Spezialfall von Modell 2, in dem der Regressionskoeffizient der Vorwissensvariable (MK3 im Fachbereich Mathematik bzw. DK3 im Fachbereich Deutsch) auf 0 fixiert ist.

Die inferenzstatistische Prüfung der in Kapitel 5 postulierten Hypothesen erfolgt mittels eines R^2-Differenzentests. Dazu werden die Determinationskoffizienten aus je zwei genesteten Modellen A und B auf signifikante Unterschiede untersucht, um zu prüfen, ob eine sparsamere Parametrisierung ausreichend ist. Die zugehörige statistische Nullhypothese lautet:

$$H_0 : \; R^2_A - R^2_B = 0 \tag{7.2}$$

Dabei ist R^2_A der Determinationskoeffizient eines komplexeren Modells A und R^2_B der Determinationskoeffizient eines jeweils restriktiveren Modells B mit weniger Parametern. Der Signifikanztest erfolgt über die Teststatistik (z. B. Steyer, 2003, S. 160):

$$F = \frac{(R^2_A - R^2_B) \,/\, (n_A - n_B)}{(1 - R^2_A) \,/\, (N - n_A)} \,, \tag{7.3}$$

wobei n_A die Anzahl der Parameter im komplexeren Modell A, n_B die Anzahl der Parameter im eingeschränkteren Modell B und N die Stichprobengröße ist. Diese Teststatistik ist F-verteilt mit $df_1 = n_A - n_B$ Zählerfreiheitsgraden und $df_2 = N - n_A$ Nennerfreiheitsgraden.

Tabelle 7.5: Determinationskoeffizient $R^2_{Y|\mathbf{Z}}$ pro Modell im Fach Mathematik (MK8)

Modellselektion	Kovariatenselektion						
	CAM	VAM			CVA		
bedingt lineare Parametrisierung (inkl. Interaktionen)	$M1^a$	$M2$	$M3$	$M4$	$M5$	$M6$	$M7$
	.499	.595	.702	.719	.623	.722	.748
lineare Parametrisierung (ohne Interaktionen)	$M8$	$M9$	$M10$	$M11$	$M12$	$M13$	$M14$
	.458	.561	.680	.692	.571	.690	.701

Anmerkungen. CAM = Contextualized Attainment Model, VAM = Value-Added Model, CVA = Contextual Value-Added Model, M = Modell.
[a] Saturiertes Zellenmittelwertemodell.

Determinationskoeffizient $R^2_{Y|\mathbf{Z}}$ im Fach Mathematik

Tabelle 7.5 enthält die Determinationskoeffizienten $R^2_{Y|\mathbf{Z}}$ der 14 Modelle für die Testleistung im Fach Mathematik (MK8). Zusätzlich veranschaulicht Abbildung 7.11 die Ergebnisse aus Tabelle 7.5 grafisch mittels eines Balkendiagramms. Die Balken repräsentieren jeweils den Prozentsatz erklärter Varianz an der Gesamtvarianz der Testwertvariable im Fach Mathematik.

Bedingt lineare Parametrisierung (mit Interaktionen). Betrachten wir zunächst die linke Seite in Abbildung 7.11, die den Prozentsatz erklärter Varianz der Modelle 1 bis 7 (saturierte und bedingt lineare Parametrisierung) in der Farbe Cyan wiedergibt. Die Tabelle 7.6 (obere Hälfte) zeigt die Ergebnisse der zugehörigen R^2-Differenzentests.

(1) *CAM vs. VAM*:
Der Anteil erklärter Varianz an der Gesamtvarianz der Mathematikleistung MK8 in Modell 1 beträgt 49.91%. Durch die Hinzunahme des fachspezifischen Vorwissens aus Klassenstufe 3 (MK3) in das Modell – zusätzlich zu den Kovariaten in Modell 1 – steigt der Anteil erklärter Varianz um 10% (von 49.91% in Modell 1 auf 59.47% in Modell 2). Wird anstatt des fachspezifischen Vorwissens

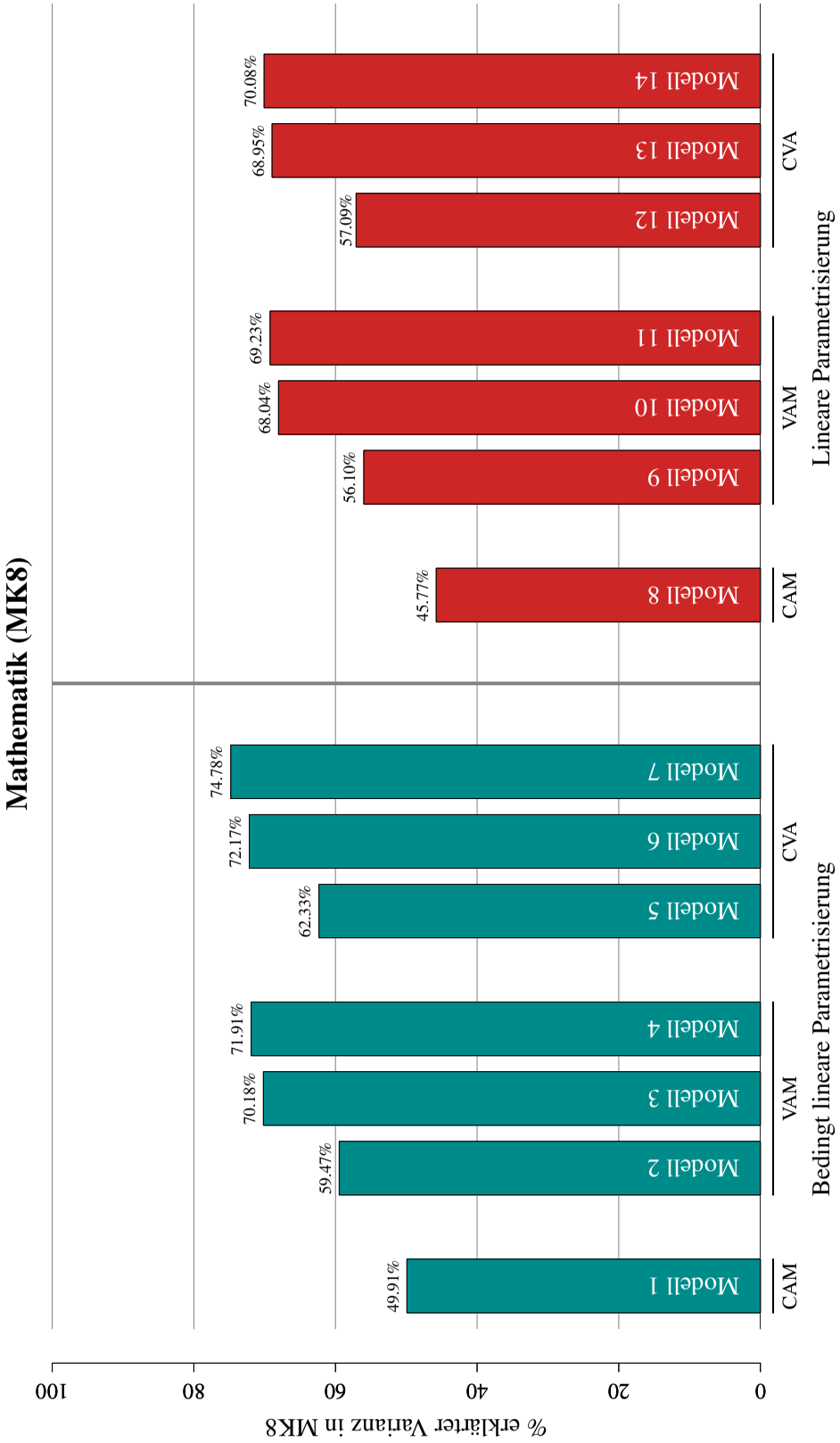

Abbildung 7.11: Prozentsatz erklärter Varianz an der Gesamtvarianz der Mathematikleistung in Klassenstufe 8 (MK8)

aus Klassenstufe 3 (MK3) das fachspezifische Vorwissen aus Klassenstufe 6 (MK6) in das Modell aufgenommen, beträgt der Zuwachs sogar 20% (von 49.91% in Modell 1 auf 70.18% in Modell 3). Werden nun beide Vorwissensvariablen – sowohl MK3 als MK6 – zusätzlich in das Modell aufgenommen, beträgt der Zuwachs insgesamt 22% (von 49.91% in Modell 1 auf 71.91% in Modell 4).
Sämtliche der drei berichteten Unterschiede zwischen den Determinationskoeffizienten sind statistisch signifikant (vgl. Tabelle 7.6). Somit zeigt sich insgesamt beim Wechsel von CAM zu VAM ein signifikanter und zudem deutlicher Zuwachs im $R^2_{Y|\mathbf{Z}}$ durch die Hinzunahme des fachspezifischen Vorwissens. Dieser Befund stützt die in der Hypothese 1.1 getroffene Annahme über den Einfluss der Hinzunahme des fachspezifischen Vorwissens in das Adjustierungsmodell.

(2) *VAM vs. CVA*:
Welchen Einfluss hat die zusätzliche Berücksichtigung von Klassenkompositionsmerkmalen (d. h. Mittelwert und Standardabweichung des fachspezifischen Vorwissens aus Klasse 3, Klasse 6 bzw. Klasse 3 und 6) auf den Determinationskoeffizienten $R^2_{Y|\mathbf{Z}}$? Um diese Frage zu adressieren, vergleichen wir die Determinationskoeffizienten $R^2_{Y|\mathbf{Z}}$ folgender genesteter Modelle: Modell 2 vs. 5, Modell 3 vs. 6 und Modell 4 vs. 7.
Werden zusätzlich zu den ursprünglichen Kovariaten *und* zum fachspezifischen Vorwissen aus Klassenstufe 3 (MK3) auch die entsprechenden Klassenkompositionsmerkmale hinsichtlich des Vorwissens aus Klassenstufe 3 in das Modell aufgenommen, steigt der Anteil erklärter Varianz um ca. 3% von 59.47% (Modell 2) auf 62.33% (Modell 5). Etwas geringer ist dieser Anstieg infolge der Hinzunahme des fachspezifischen Vorwissens in Klassenstufe 6 und der entsprechenden Kompositionsmerkmale: Hier steigt der Prozentsatz erklärter Varianz um 2% von 70.18% (Modell 3) auf 72.17% (Modell 6). Stehen Informationen sowohl zum fachspezifischen Vorwissen aus Klassenstufe 3 (MK3) als auch Klassenstufe 6 (MK6) zur Verfügung, so steigt der Prozentsatz erklärter Varianz um ca. 3% von 71.91% (Modell 4) auf 74.78% (Modell 7).
Alle drei berichteten Unterschiede zwischen den Determinationskoeffizienten sind *nicht* signifikant (vgl. Tabelle 7.6). Somit spricht dieser Befund gegen die in Hypothese 1.2 getroffene Annahme über den zusätzlichen Einfluss der leistungsmäßigen Klassenkomposition – zusätzlich zum Vorwissen und den restlichen Ko-

Tabelle 7.6: R^2-Differenzen genesteter Modelle: Modifikation der Kovariatenselektion im Fach Mathematik (MK8)

Modellvergleich		ΔR^2	F-Wert	df_1	df_2	p-Wert
bedingt lineare Parametrisierung (inkl. Interaktionen):						
CAM vs. VAM	$M1^a \rightarrow M2$	.10	8.91	320	12 068	<.001
	$M1 \rightarrow M3$	.20	25.03	320	12 068	<.001
	$M1 \rightarrow M4$	.22	9.23	960	11 428	<.001
VAM vs. CVA	$M2 \rightarrow M5$	.03	0.42	1 920	10 148	.999
	$M3 \rightarrow M6$	.02	0.38	1 920	10 148	.999
	$M4 \rightarrow M7$	.03	0.24	3 840	7 588	.999
bedingte UA[b]	$M3 \rightarrow M4$	.02	1.20	640	11 428	.001
	$M6 \rightarrow M7$	.03	0.34	2 560	7 588	.999
lineare Parametrisierung (ohne Interaktionen):						
CAM vs. VAM	$M8 \rightarrow M9$	.10	2 957.21	1	12 700	<.001
	$M8 \rightarrow M10$	.22	8 691.89	1	12 700	<.001
	$M8 \rightarrow M11$	.23	3 368.03	2	12 699	<.001
VAM vs. CVA	$M9 \rightarrow M12$	.01	161.62	2	12 698	<.001
	$M10 \rightarrow M13$	.01	204.04	2	12 698	<.001
	$M11 \rightarrow M14$	.01	202.02	2	12 697	<.001
bedingte UA[b]	$M9 \rightarrow M10$	.01	533.55	1	12 699	<.001
	$M13 \rightarrow M14$	.01	529.42	1	12 697	<.001

Anmerkungen. CAM = Contextualized Attainment Model, VAM = Value-Added Model, CVA = Contextual Value-Added Model, M = Modell.

[a] Saturiertes Zellenmittelwertemodell.

[b] Bedingte Unabhängigkeit der Variable MK8 von MK3 gegeben MK6 und der weiteren Kovariaten im Adjustierungsmodell.

variaten. Da im Rahmen der vorliegenden Arbeit jedoch insbesondere die Sensitivität der Effektschätzungen individueller Klassen fokussiert wird, werde ich dies nachfolgend auch anhand weiterer Kriterien beurteilen (vgl. Abschnitt 7.3.3 bis 7.3.5).

(3) *Bedingte Unabhängigkeit* [11]:
Durch die Hinzunahme von MK6 – zusätzlich zu MK3 und den restlichen Kovariaten im Adjustierungsmodell – ist der stärkste Zuwachs im Anteil erklärter Varianz zu verzeichnen (Modell 1 vs. Modell 4). Jedoch zeigt sich ein ähnlich hoher Anstieg im Anteil erklärter Varianz, wenn allein MK6 (ohne MK3, aber gleichfalls zusätzlich zu den restlichen Kovariaten) in das Modell aufgenommen wird (Modell 1 vs. Modell 3). Somit steigt der Anteil erklärter Varianz um weniger als 2%, wenn man – anstatt MK6 allein – beide Vorwissensvariablen (MK3 *und* MK6) in das Modell aufnimmt (von 70.18% in Modell 3 auf 71.91% in Modell 4). Hier stellt sich die Frage, ob man auf die Hinzunahme des fachspezifischen Vorwissens aus Klassenstufe 3 (MK3) verzichten kann, wenn man bereits das fachspezifische Vorwissen in Klassenstufe 6 (MK6) in das Modell aufgenommen hat? Aus theoretischer Sicht ist dies dann zutreffend, wenn die zusätzliche Berücksichtigung von MK3 über MK6 hinaus keine weitere Veränderung im $R^2_{Y|\mathbf{Z}}$ verursacht. Dies ist bspw. dann der Fall, wenn die Testwertvariable MK8 bedingt regressiv unabhängig von MK3 gegeben MK6 und der weiteren Kovariaten im Modell ist (*bedingte regressive Unabhängigkeit von* MK3). Betrachtet man lediglich die Effektstärke ΔR^2, ist dies m. E. vertretbar – nicht zuletzt aus testökonomischen Gründen –, denn der Anteil erklärter Varianz verändert sich um weniger als 2%. Jedoch wird diese Differenz der Determinationskoeffizienten statistisch signifikant, $F(640, 11\,428) = 1.20$, $p = .001$ (vgl. Tabelle 7.6). Dieses Ergebnis spricht gegen die Annahme bedingter Unabhängigkeit.
Im Gegensatz dazu ist der Unterschied der Determinationskoeffizienten zwischen Modell 6 und Modell 7 nicht signifikant, $F(2\,560, 7\,588) = 0.34$, $p = .999$ (vgl.

[11] Bedingte Unabhängigkeit bezeichnet nachfolgend – je nach Modellvergleich – die bedingte Unabhängigkeit der Testwertvariablen MK8 von MK3 gegeben MK6 und der weiteren Kovariaten (Modell 3 vs. 4, Modell 9 vs. 10) bzw. die bedingte Unabhängigkeit der Testwertvariablen MK8 von MK3 und der entsprechenden leistungsmäßigen Klassenkomposition gegeben MK6, der leistungsmäßigen Klassenkomposition bezüglich MK6 und der weiteren Kovariaten (Modell 6 vs. 7, Modell 13 vs. 14). Zum Zwecke der Übersichtlichkeit verwende ich für Aussagen über eben diese Modellvergleiche nachfolgend stets die Kurzform *bedingte Unabhängigkeit*.

Tabelle 7.6). Dieser Befund spricht somit für die entsprechende Annahme der bedingten Unabhängigkeit von MK3 und der auf MK3 basierenden leistungsmäßigen Klassenkomposition.

Lineare Parametrisierung (ohne Interaktionen). Betrachten wir nun die rechte Seite in Abbildung 7.11. Hier ist der Prozentsatz erklärter Varianz der Modelle 8 bis 14 (lineare Parametrisierung ohne Interaktionen) in Rot dargestellt. Die untere Hälfte von Tabelle 7.6 zeigt die zugehörigen Ergebnisse der R^2-Differenzentests.

(1) *CAM vs. VAM*:
Insgesamt zeigt sich auch für die lineare Parametrisierung ein deutlicher Zuwachs im $R^2_{Y|Z}$ durch die Hinzunahme des fachspezifischen Vorwissens. Auch hier ist der stärkste Zuwachs im Anteil erklärter Varianz durch die Hinzunahme des fachspezifischen Vorwissens aus Klassenstufe 6 (MK6) zu verzeichnen. Dieser steigt hier sogar um ca. 22% von 45.77% (Modell 8) auf 68.04% (Modell 10) bzw. um 23% von 45.77% (Modell 8) auf 69.23% (Modell 11). Zudem sind wiederum alle drei Unterschiede zwischen den Determinationskoeffizienten statistisch signifikant (vgl. Tabelle 7.6). Somit ist auch dieser Befund hypothesenkonform.

(2) *VAM vs. CVA*:
Werden nun zusätzlich die entsprechenden Klassenkompositionsmerkmale in die Modelle aufgenommen, steigt auch hier der Anteil erklärter Varianz an. Dieser Anstieg schwankt zwischen 0.85% (Modell 11 vs. Modell 14) und 0.99% (Modell 9 vs. Modell 12). Der Zuwachs im $R^2_{Y|Z}$ durch die zusätzliche Modellierung der leistungsmäßigen Klassenkomposition ist somit insgesamt geringer als bei der bedingt linearen Parametrisierung, bei der auch Interaktionen zwischen den Kovariaten modelliert werden. Jedoch sind diese Unterschiede zwischen den Determinationskoeffizienten aus den linearen Modellen statistisch signifikant (vgl. Tabelle 7.6).

(3) *Bedingte Unabhängigkeit*:
Auch die Unterschiede zwischen Modell 9 und 10 bzw. zwischen Modell 13 und 14 sind marginal geringer als bei den entsprechenden bedingt linearen Modellen. Allerdings werden beide Differenzen der Determinationskoeffizienten statistisch signifikant (vgl. Tabelle 7.6). Dies spricht gegen die Plausibilität der zugehörigen bedingten Unabhängigkeitsannahmen.

Tabelle 7.7: R^2-Differenzen genesteter Modelle: Modifikation der Parametrisierung im Fach Mathematik (MK8)

Modellvergleich		ΔR^2	F-Wert	df_1	df_2	p-Wert
CAM	$M1^a \rightarrow M8$	.04	3.34	313	12 388	<.001
VAM	$M2 \rightarrow M9$	.04	1.67	632	12 068	<.001
	$M3 \rightarrow M10$	.02	1.36	632	12 068	<.001
	$M4 \rightarrow M11$	.03	0.87	1 271	11 428	.999
CVA	$M5 \rightarrow M12$	.05	0.57	2 550	10 148	.999
	$M6 \rightarrow M13$	.03	0.45	2 550	10 148	.999
	$M7 \rightarrow M14$	.05	0.28	5 109	7 588	.999

Anmerkungen. CAM = Contextualized Attainment Model, VAM = Value-Added Model, CVA = Contextual Value-Added Model, M = Modell.
[a] Saturiertes Zellenmittelwertemodell.

Bedingt lineare vs. lineare Parametrisierung. Vergleicht man die Modelle *innerhalb* einer Parametrisierungsform, so zeigt sich, dass das Muster der Veränderung im $R^2_{Y|\mathbf{Z}}$ infolge der Modifikation der Kovariatenselektion ähnlich ist – unabhängig davon, ob man die komplexere (bedingt lineare) Parametrisierung oder die lineare Parametrisierung ohne Interaktionen wählt: Je mehr Kovariaten im Adjustierungsmodell enthalten sind, desto größer der Anteil erklärter Varianz und desto geringer die Zuwächse im Anteil erklärter Varianz infolge Hinzunahme weiterer Kovariaten.

Vergleicht man nun *zwischen* den beiden Parametrisierungsformen – d. h. zwischen einander hinsichtlich der Kovariatenselektion entsprechenden Modellen, die sich lediglich in der Parametrisierung (Modellselektion) unterscheiden –, fällt der recht stabile Unterschied im $R^2_{Y|\mathbf{Z}}$ auf: So beträgt der Anteil erklärter Varianz 45.77% in Modell 8, während dieser in Modell 1 bei 49.91% liegt. Derartige Unterschiede, die allein auf die Modellselektion und nicht auf die Wahl der Kovariaten zurückzuführen sind, lassen sich bei allen sieben paarweisen Vergleichen der Modelle mit jeweils identischer Kovariatenselektion finden. Die Unterschiede im Anteil erklärter Varianz schwanken zwischen minimal 2% (Modell 3 vs. Modell 10) und maximal 5% (Modell 5 vs. Modell 12). Dabei wird der Unterschied im Anteil erklärter Varianz zwar geringer, wenn das fachspezifische Vorwissen in das Adjustierungsmodell aufgenommen wird (d. h. beim Wechsel von CAM zum VAM). Jedoch nimmt dieser Unterschied wieder zu, wenn zu-

sätzlich auch die leistungsmäßige Klassenkomposition (CVA) modelliert wird.

Bei den zugehörigen R^2-Differenzentests (vgl. Tabelle 7.7) sind lediglich die ersten drei Modellunterschiede statistisch signifikant (Modell 1 vs. 8, Modell 2 vs. 9 und Modell 3 vs. 10, vgl. Tabelle 7.7). Sind beide Vorwissensvariablen MK3 und MK6 zusätzlich im Adjustierungsmodell enthalten (Modell 4 vs. 11), sind die R^2-Differenzen nicht signifikant. Und auch die restlichen Unterschiede zwischen den drei Modellen des Typs CVA sind nicht signifikant. Insgesamt stützt diese Befundlage einerseits Hypothese 2, dass die komplexere (bedingt lineare) Parametrisierung der linearen Parametrisierung (ohne Interaktionen) vorzuziehen ist. Andererseits stützen die Ergebnisse zusätzlich auch die Annahme einer Interaktion zwischen Kovariaten- und Modellselektion (Hypothese 3). So sind die Unterschiede zwischen den Determinationskoeffizienten zwischen einander hinsichtlich der Kovariaten entsprechenden Modellen nicht mehr signifikant, sobald neben den weiteren Kovariaten auch beide Vorwissensvariablen enthalten sind.

Determinationskoeffizient $R^2_{Y|\mathbf{Z}}$ im Fach Deutsch

Tabelle 7.8 enthält die Determinationskoeffizienten $R^2_{Y|\mathbf{Z}}$ der 14 Modelle für die Testleistung im Fach Deutsch (DK8). Abbildung 7.12 zeigt diese Ergebnisse grafisch mittels eines Balkendiagramms. Die Balken repräsentieren jeweils den Prozentsatz erklärter Varianz an der Gesamtvarianz der Testwertvariable im Fach Deutsch.

Bedingt lineare Parametrisierung (mit Interaktionen). Auch für die Testleistung im Fach Deutsch betrachten wir zunächst die Ergebnisse der saturierten und bedingt linearen Parametrisierung mit Interaktionen (Modell 1 bis 7), die auf der linken Seite des Balkendiagramms in Abbildung 7.12 wiedergegeben sind. Die obere Hälfte in Tabelle 7.9 zeigt die Ergebnisse der entsprechenden R^2-Differenzentests.

(1) *CAM vs. VAM*:
Der Anteil erklärter Varianz an der Gesamtvarianz der Deutschleistung DK8 in Modell 1 beträgt 41.68%. Durch die Hinzunahme des fachspezifischen Vorwissens aus Klassenstufe 3 (DK3) in das Adjustierungsmodell – wiederum zusätzlich zu den Kovariaten in Modell 1 – steigt der Anteil erklärter Varianz um 11% auf 52.86% in Modell 2. Wird anstatt das fachspezifische Vorwissen aus Klassenstufe 3 (DK3) das fachspezifische Vorwissen aus Klassenstufe 6 (DK6) in das Modell aufgenommen, beträgt der Zuwachs hingegen knapp 15% (56.45% in Modell 3).

Tabelle 7.8: Determinationskoeffizient $R^2_{Y|\mathbf{Z}}$ pro Modell im Fach Deutsch (DK8)

Modellselektion	Kovariatenselektion						
	CAM	VAM			CVA		
bedingt lineare Parametrisierung (inkl. Interaktionen)	$M1^a$ .417	$M2$ .529	$M3$ .565	$M4$ .602	$M5$ .565	$M6$ .607	$M7$ .654
lineare Parametrisierung (ohne Interaktionen)	$M8$ .379	$M9$ .497	$M10$ .540	$M11$ .569	$M12$ .505	$M13$ .541	$M14$ .572

Anmerkungen. CAM = Contextualized Attainment Model, VAM = Value-Added Model, CVA = Contextual Value-Added Model, M = Modell.
[a] Saturiertes Zellenmittelwertemodell.

Werden beide Vorwissensvariablen – sowohl das fachspezifische Vorwissen aus Klassenstufe 3 (DK3) als auch aus Klassenstufe 6 (DK6) – zusätzlich in das Modell aufgenommen, beträgt der Zuwachs ca. 18% (von 41.68% in Modell 1 auf 60.20% in Modell 4).

Sämtliche der drei berichteten Unterschiede zwischen den Determinationskoeffizienten sind statistisch signifikant (vgl. Tabelle 7.9). Somit zeigt sich beim Wechsel von CAM zu VAM – wie bereits im Fachbereich Mathematik – ebenso für das Fach Deutsch ein signifikanter Zuwachs im $R^2_{Y|\mathbf{Z}}$ durch die Hinzunahme des fachspezifischen Vorwissens. Dieser Befund stützt die in Hypothese 1.1 getroffene Annahme über den Einfluss der Hinzunahme des fachspezifischen Vorwissens in das Adjustierungsmodell.

(2) *VAM vs. CVA*:

Um den Einfluss der Klassenkompositionsmerkmale auf den Anteil erklärter Varianz zu quantifizieren, vergleichen wir wiederum paarweise den Determinationskoeffizienten $R^2_{Y|\mathbf{Z}}$ jeweils folgender genesteter Modelle: Modell 2 vs. 5, Modell 3 vs. 6 und Modell 4 vs. 7.

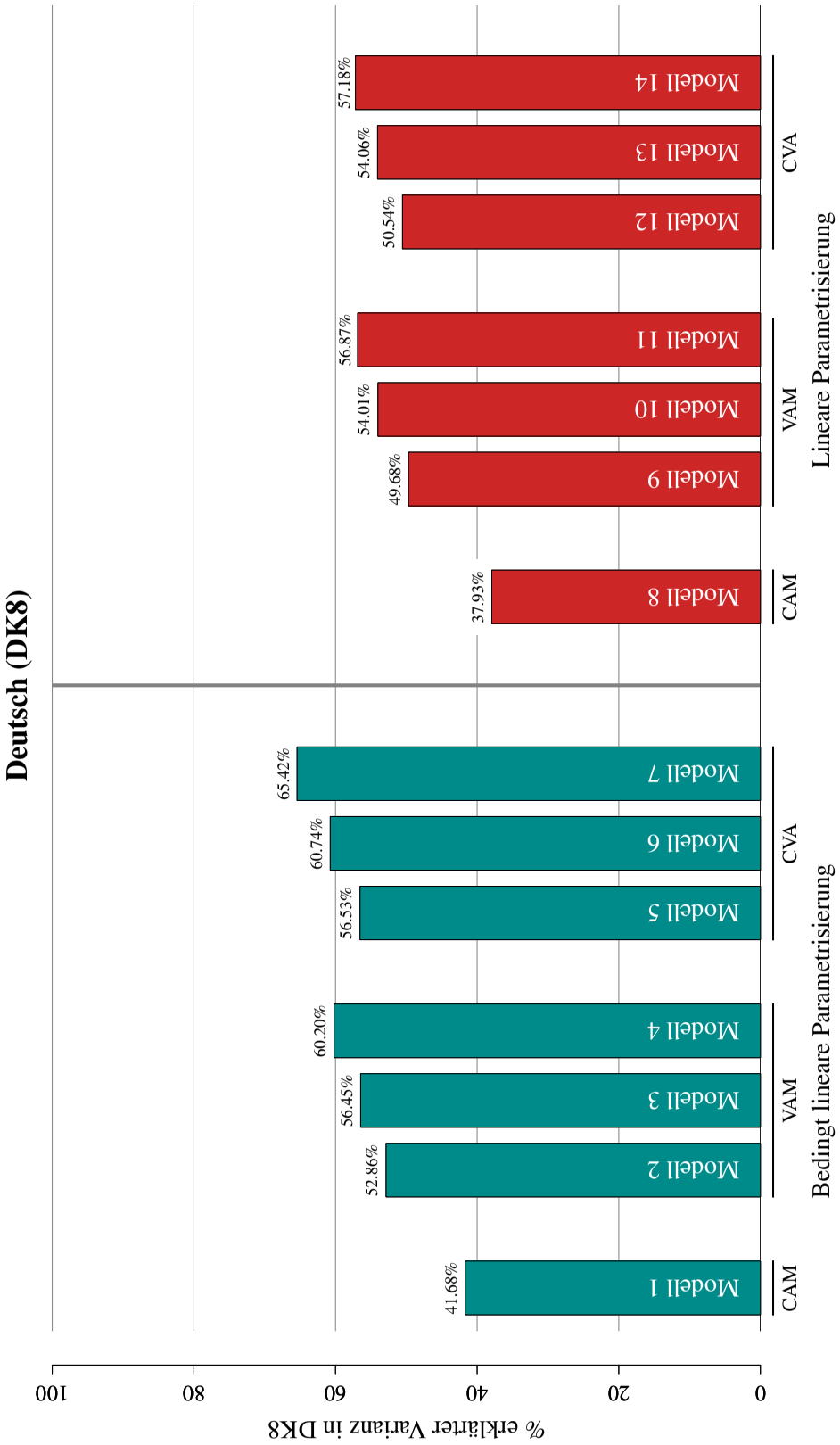

Abbildung 7.12: Prozentsatz erklärter Varianz an der Gesamtvarianz der Deutschleistung in Klassenstufe 8 (DK8)

Werden zusätzlich zu den ursprünglichen Kovariaten *und* zum fachspezifischen Vorwissen aus Klassenstufe 3 (DK3) auch die entsprechenden Klassenkompositionsmerkmale hinsichtlich des Vorwissens aus Klassenstufe 3 in das Modell aufgenommen, steigt der Anteil erklärter Varianz um ca. 4% von 52.86% (Modell 2) auf 56.53% (Modell 5). Dieser Zuwachs im $R^2_{Y|\mathbf{Z}}$ ist vergleichbar für das fachspezifische Vorwissen in Klassenstufe 6 und die entsprechenden Kompositionsmerkmale: Auch hier steigt der Prozentsatz erklärter Varianz um ca. 4% von 56.45% (Modell 3) auf 60.74% (Modell 6). Stehen Informationen sowohl zum fachspezifischen Vorwissen aus Klassenstufe 3 (DK3) als auch Klassenstufe 6 (DK6) zur Verfügung, so steigt der Prozentsatz erklärter Varianz um 5% von 60.20% (Modell 4) auf 65.42% (Modell 7).

Wie bereits im Fach Mathematik sind sämtliche der drei berichteten Unterschiede zwischen den Determinationskoeffizienten nicht signifikant (vgl. Tabelle 7.9). Somit spricht dieser Befund auch im Fachbereich Deutsch gegen die in Hypothese 1.2 getroffene Annahme über den zusätzlichen Einfluss der leistungsmäßigen Klassenkomposition – zusätzlich zum Vorwissen und den restlichen Kovariaten. Da im Rahmen der vorliegenden Arbeit jedoch insbesondere die Sensitivität der Effektschätzungen individueller Klassen fokussiert wird, werden nachfolgend zusätzlich weitere Kriterien herangezogen (vgl. Abschnitt 7.3.3 bis 7.3.5).

(3) *Bedingte Unabhängigkeit*:

Auch im Fachbereich Deutsch ist durch die Hinzunahme von DK6 zusätzlich zu DK3 und den restlichen Kovariaten im Adjustierungsmodell der stärkste Zuwachs im Anteil erklärter Varianz zu verzeichnen (Modell 1 vs. Modell 4). Im Unterschied zum Fachbereich Mathematik zeigt sich hier ein deutlicherer Zugewinn im Anteil erklärter Varianz, wenn man – anstelle des fachspezifischen Vorwissens in Klassenstufe 6 allein – beide Vorwissensvariablen (DK3 und DK6) berücksichtigt (von 56.45% in Modell 3 auf 60.20% in Modell 4). Diese Differenz der Determinationskoeffizienten ist statistisch signifikant, $F(640, 11\,428) = 1.72$, $p < .001$ (vgl. Tabelle 7.9). Dieses Ergebnis spricht somit auch hier gegen die Annahme bedingter Unabhängigkeit der Testwertvariable DK8 von DK3 gegeben DK6 und der weiteren Kovariaten im Modell.

Im Gegensatz dazu ist der Unterschied der Determinationskoeffizienten zwischen Modell 6 und Modell 7 nicht signifikant, $F(2\,560, 7\,588) = 0.41$, $p = .999$ (vgl.

Tabelle 7.9: R^2-Differenzen genesteter Modelle: Modifikation der Kovariatenselektion im Fach Deutsch (DK8)

Modellvergleich		ΔR^2	F-Wert	df_1	df_2	p-Wert
bedingt lineare Parametrisierung (inkl. Interaktionen):						
CAM vs. VAM	$M1^a \rightarrow M2$	.11	8.96	320	12 068	<.001
	$M1 \rightarrow M3$	.15	12.66	320	12 068	<.001
	$M1 \rightarrow M4$	.19	5.53	960	11 428	<.001
VAM vs. CVA	$M2 \rightarrow M5$	.04	0.46	1 920	10 148	.999
	$M3 \rightarrow M6$	.04	0.58	1 920	10 148	.999
	$M4 \rightarrow M7$	.05	0.30	3 840	7 588	.999
bedingte UA[b]	$M3 \rightarrow M4$	.04	1.72	640	11 428	<.001
	$M6 \rightarrow M7$	.05	0.41	2 560	7 588	.999
lineare Parametrisierung (ohne Interaktionen):						
CAM vs. VAM	$M8 \rightarrow M9$	.12	2 959.13	1	12 700	<.001
	$M8 \rightarrow M10$	.16	4 403.37	1	12 700	<.001
	$M8 \rightarrow M11$	.19	2 380.99	2	12 699	<.001
VAM vs. CVA	$M9 \rightarrow M12$	.01	103.25	2	12 698	<.001
	$M10 \rightarrow M13$	<.01	7.96	2	12 698	<.001
	$M11 \rightarrow M14$	<.01	40.39	2	12 697	<.001
bedingte UA[b]	$M9 \rightarrow M10$	.03	851.70	1	12 699	<.001
	$M13 \rightarrow M14$	.03	920.70	1	12 697	<.001

Anmerkungen. CAM = Contextualized Attainment Model, VAM = Value-Added Model, CVA = Contextual Value-Added Model, M = Modell.

[a] Saturiertes Zellenmittelwertemodell.

[b] Bedingte Unabhängigkeit der Variable DK8 von DK3 gegeben DK6 und der weiteren Kovariaten im Adjustierungsmodell.

Tabelle 7.9). Dieser Befund spricht somit für die entsprechende Annahme der bedingten Unabhängigkeit von DK3 und der auf DK3 basierenden leistungsmäßigen Klassenkomposition. Zudem ist das Ergebnis konkordant mit den Ergebnissen im Fach Mathematik.

Lineare Parametrisierung (ohne Interaktionen). Kommen wir nun zur rechten Seite in Abbildung 7.12. Hier ist – wiederum in der Farbe Rot – der Prozentsatz erklärter Varianz der Modelle 8 bis 14 (lineare Parametrisierung ohne Interaktionen) dargestellt. Die untere Hälfte von Tabelle 7.9 zeigt zudem die zugehörigen Ergebnisse der R^2-Differenzentests.

(1) *CAM vs. VAM*:
Insgesamt zeigt sich gleichsam für die lineare Parametrisierung ein deutlicher Zuwachs in $R^2_{Y|\mathbf{Z}}$ durch die Hinzunahme des fachspezifischen Vorwissens. Auch hier ist der stärkste Zuwachs im Anteil erklärter Varianz durch die Hinzunahme des fachspezifischen Vorwissens aus Klassenstufe 6 (DK6) zu verzeichnen. Dieser steigt um ca. 16% von 37.93% (Modell 8) auf 54.01% (Modell 10) bzw. um ca. 19% von 37.93% (Modell 8) auf 56.87% (Modell 11). Zudem sind wiederum sämtliche der drei Unterschiede zwischen den Determinationskoeffizienten statistisch signifikant (vgl. Tabelle 7.9). Somit ist auch dieser Befund hypothesenkonform.

(2) *VAM vs. CVA*:
Werden nun zusätzlich Klassenkompositionsmerkmale in die Modelle aufgenommen, steigt auch im Rahmen der linearen Parametrisierung im Fach Deutsch der Anteil erklärter Varianz an. Dieser Anstieg schwankt zwischen minimal 0.05% (Modell 10 vs. Modell 13) und maximal 0.86% (Modell 9 vs. Modell 12). Der Zuwachs im $R^2_{Y|\mathbf{Z}}$ durch die zusätzliche Modellierung der leistungsmäßigen Klassenkomposition ist somit insgesamt geringer – sowohl im Vergleich zu der Veränderung bei bedingt linearer Parametrisierung im Fachbereich Deutsch als auch im Vergleich zu den entsprechenden Veränderungen im Fachbereich Mathematik. Jedoch sind auch diese Unterschiede zwischen den Determinationskoeffizienten aus den linearen Modellen statistisch signifikant (vgl. Tabelle 7.9).

(3) *Bedingte Unabhängigkeit*:
Auch die Unterschiede hinsichtlich der Varianzaufklärung zwischen Modell 9

Tabelle 7.10: R^2-Differenzen genesteter Modelle: Modifikation der Parametrisierung im Fach Deutsch (DK8)

Modellvergleich		ΔR^2	F-Wert	df_1	df_2	p-Wert
CAM	$M1^a \rightarrow M8$	.04	2.51	313	12 388	<.001
VAM	$M2 \rightarrow M9$	.03	1.28	632	12 068	<.001
	$M3 \rightarrow M10$	.02	1.04	632	12 068	.225
	$M4 \rightarrow M11$	.03	0.75	1 271	11 428	.999
CVA	$M5 \rightarrow M12$	.06	0.56	2 550	10 148	.999
	$M6 \rightarrow M13$	.07	0.67	2 550	10 148	.999
	$M7 \rightarrow M14$	.08	0.36	5 109	7 588	.999

Anmerkungen. CAM = Contextualized Attainment Model, VAM = Value-Added Model, CVA = Contextual Value-Added Model, M = Modell.
[a] Saturiertes Zellenmittelwertemodell.

und 10 bzw. zwischen Modell 13 und 14 sind etwas kleiner als bei den entsprechenden bedingt linearen Modellen. Allerdings werden beide Differenzen der Determinationskoeffizienten signifikant (vgl. Tabelle 7.9). Dies spricht gegen die Plausibilität der entsprechenden bedingten Unabhängigkeitsannahmen.

Bedingt lineare vs. lineare Parametrisierung. Vergleicht man die Modelle *innerhalb* einer Parametrisierungsform, so zeigt sich, dass auch im Fachbereich Deutsch das Muster der Veränderung im $R^2_{Y|\mathbf{Z}}$ infolge der Modifikation der Kovariatenselektion ähnlich ist – unabhängig davon, ob man die komplexere (bedingt lineare) Parametrisierung oder die lineare Parametrisierung ohne Interaktionen wählt: Je mehr Kovariaten im Adjustierungsmodell enthalten sind, desto größer der Anteil erklärter Varianz und desto geringer die Zuwächse im Anteil erklärter Varianz infolge Hinzunahme weiterer Kovariaten.

Vergleicht man nun *zwischen* den beiden Parametrisierungsformen – d. h. zwischen einander hinsichtlich der Kovariatenselektion entsprechenden Modellen, die sich lediglich in der Parametrisierung (Modellselektion) unterscheiden –, fällt auch hier der recht stabile Unterschied im $R^2_{Y|\mathbf{Z}}$ auf: So beträgt der Anteil erklärter Varianz 37.93% in Modell 8, während dieser in Modell 1 bei 41.68% liegt. Solche Unterschiede, die allein auf die Modellselektion und nicht auf die Wahl der Kovariaten zurückzuführen

sind, lassen sich auch im Fach Deutsch bei allen sieben paarweisen Vergleichen der Modelle mit jeweils identischer Kovariatenselektion finden. Die Unterschiede im Anteil erklärter Varianz schwanken zwischen minimal 2% (Modell 3 vs. Modell 10) und maximal 8% (Modell 7 vs. Modell 14). Dabei wird der Unterschied im Anteil erklärter Varianz beim paarweisen Vergleich einander hinsichtlich der Kovariatenselektion entsprechenden Modellen zwar geringer, wenn das fachspezifische Vorwissen in das Adjustierungsmodell aufgenommen wird. Jedoch nimmt dieser Unterschied wiederum zu, wenn zusätzlich auch die leistungsmäßige Klassenkomposition modelliert wird.

Bei den zugehörigen R^2-Differenzentests (vgl. Tabelle 7.10) sind lediglich die ersten zwei Modellunterschiede statistisch signifikant (Modell 1 vs. 8 und Modell 2 vs. 9). Ist das fachspezifische Vorwissen MK6 zusätzlich im Adjustierungsmodell enthalten (Modell 3 vs. 10), so sind die R^2-Differenzen nicht signifikant. Und auch die restlichen Unterschiede zwischen den drei Modellen des Typs CVA sind wiederum nicht signifikant. Insgesamt stützt diese Befundlage einerseits Hypothese 2, dass die komplexere (bedingt lineare) Parametrisierung der linearen Parametrisierung (ohne Interaktionen) vorzuziehen ist. Andererseits stützen die Ergebnisse zusätzlich auch die Annahme einer Interaktion zwischen Kovariaten- und Modellselektion (Hypothese 3). So sind die Unterschiede zwischen den Determinationskoeffizienten zwischen einander hinsichtlich der Kovariaten entsprechenden Modellen nicht mehr signifikant, sobald neben den weiteren Kovariaten auch das fachspezifische Vorwissen aus Klassenstufe 6 enthalten sind.

Zusammenfassung: Determinationskoeffizient $R^2_{Y|Z}$

Zusammenfassend lässt sich für die Determinationskoeffizienten $R^2_{Y|Z}$ – erneut für beide Fachbereiche Mathematik und Deutsch – feststellen: Je komplexer das Modell, d. h. je mehr relevante Kovariaten und Parameter im Modell enthalten sind, desto höher ist der Anteil erklärter Varianz an der Gesamtvarianz der Mathematikleistung (MK8) respektive der Deutschleistung (DK8).

Innerhalb einer Parametrisierungsform zeigt sich jeweils, dass der Anteil erklärter Varianz umso höher ist, je mehr Kovariaten in das Modell aufgenommen werden. Bezüglich der Kovariatenselektion ist insbesondere die Bedeutung des fachspezifischen Vorwissens in Klassenstufe 6 zu nennen: Die zusätzliche Berücksichtigung der Vorwissensvariable MK6 bzw. DK6 trägt zu einer signifikanten und substanziellen Erhöhung

der Varianzaufklärung bei (Hypothese 1.1). Hingegen muss Hypothese 1.2 verworfen werden: Zwar werden die R^2-Differenzen beim Wechsel vom VAM zum CVA bei Anwendung eines sparsameren linearen Adjustierungsmodells (ohne Interaktionen) statistisch signifikant. Jedoch fällt der Zugewinn hinsichtlich des Kriteriums der Varianzaufklärung infolge der zusätzlichen Berücksichtigung der leistungsmäßigen Klassenkomposition bei bedingt linearer Parametrisierung nicht signifikant aus, was gegen die Plausibilität von Hypothese 1.2 spricht. Unabhängig von der Parametrisierung ist der Zuwachs hinsichtlich des Anteils erklärter Varianz hier zudem deutlich kleiner als beim Wechsel von CAM zum VAM.

Des Weiteren zeigen sich Unterschiede infolge des Wechsels der Parametrisierung (Modellselektion): In beiden Fachbereichen sprechen die Ergebnisse dafür, dass die bedingt lineare Parametrisierung mit Berücksichtigung potenzieller Interaktionen zwischen den Kovariaten einer sparsameren linearen Parametrisierung vorzuziehen ist (Hypothese 2). Ferner kann auch Hypothese 3 auf Basis des Kriteriums der Varianzaufklärung nicht falsifiziert werden: Sobald das fachspezifische Vorwissen MK3 und MK6 (im Fachbereich Mathematik) bzw. DK6 (im Fachbereich Deutsch) zusätzlich im Adjustierungsmodell enthalten ist, sind die Unterschiede im Determinationskoeffizienten zwischen Modellen mit linearer vs. mit bedingt linearer Parametrisierung nicht mehr signifikant.

Schließlich stützt die Befundlage Hypothese 4 zur Generalisierung über die beiden Fachbereiche: Hinsichtlich des Kriteriums der Varianzaufklärung besteht keine Fachspezifität, da sich die dargelegten Ergebnismuster gleichermaßen sowohl im Fach Mathematik als auch im Fach Deutsch finden lassen.

7.3.3 Korrelationen

Nachfolgend werden die Korrelationen der adjustierten klassenspezifischen Effektschätzungen beim paarweisen Modellvergleich berichtet und einer vergleichenden Analyse unterzogen. Als quantitatives Maß des Zusammenhangs wird der Korrelationskoeffizient nach Spearman verwendet. Die Korrelation ist ein Maß für den linearen Zusammenhang zweier Variablen. In der vorliegenden Anwendung handelt es sich bei diesen Variablen jeweils um die aus den verschiedenen Modellen resultierenden adjustierten Effektschätzungen. Hohe Korrelationen indizieren eine hohe Stabilität in der Rangreihung aller Klassen basierend auf den Effektschätzungen aus zwei unterschiedlichen

Adjustierungsmodellen. Hingegen weisen niedrige Korrelationen auf eine geringe Stabilität dieser Rangreihe hin und somit auf eine hohe Sensitivität der adjustierten klassenspezifischen Effektschätzungen gegenüber der Wahl des Adjustierungsmodells. In diesem Fall würde die Modellwahl maßgeblich beeinflussen, ob der adjustierte Effekt einer Klasse – relativ zu den anderen Klassen – hoch oder niedrig bzw. positiv oder negativ ist.

Die Korrelationen sind sowohl für den Fachbereich Mathematik als auch Deutsch in den Abbildungen 7.13 und 7.14 jeweils in der oberen Dreiecksmatrix dargestellt. Die untere Dreiecksmatrix zeigt zudem die Streudiagramme, welche die adjustierten klassenspezifischen Effektschätzungen aus jeweils zwei Modellen darstellen.

Jede der beiden Korrelationsmatrizen lässt sich in drei Bereiche aufteilen: (a) Die Korrelationen zwischen den Modellen 1 bis 7 mit saturierter und bedingt linearer Parametrisierung (inkl. Interaktionen) finden sich im linken oberen Bereich der Matrix. (b) Die Korrelationen zwischen den Modellen 8 bis 14 mit linearer Parametrisierung ohne Interaktionen befinden sich im rechten unteren Bereich der Matrix. (c) Schließlich finden sich sämtliche paarweise Korrelationen zwischen Modellen mit unterschiedlicher Parametrisierung im rechten oberen (numerisch) sowie im linken unteren (grafisch) Bereich der Korrelationsmatrix. Entsprechend dieser drei Bereiche werden nachfolgend die Ergebnisse berichtet – zunächst für das Fach Mathematik und anschließend für das Fach Deutsch.

Korrelationen im Fach Mathematik

Wir betrachten zunächst die Korrelationen der klassenspezifischen Effektschätzungen im Fach Mathematik, welche in Abbildung 7.13 dargestellt sind. Insgesamt sind die Korrelationen überwiegend recht hoch, weisen jedoch auch deutliche Unterschiede auf.

Bedingt lineare Parametrisierung (mit Interaktionen). Hinsichtlich der Modelle 1 bis 7 mit saturierter und bedingt linearer Parametrisierung, in denen potenzielle Interaktionen berücksichtigt werden, zeigen sich die folgenden Korrelationen zwischen den resultierenden adjustierten Effektschätzungen:

(1) *CAM vs. VAM*:
Die korrelativen Zusammenhänge zwischen den Effektschätzungen aus dem CAM

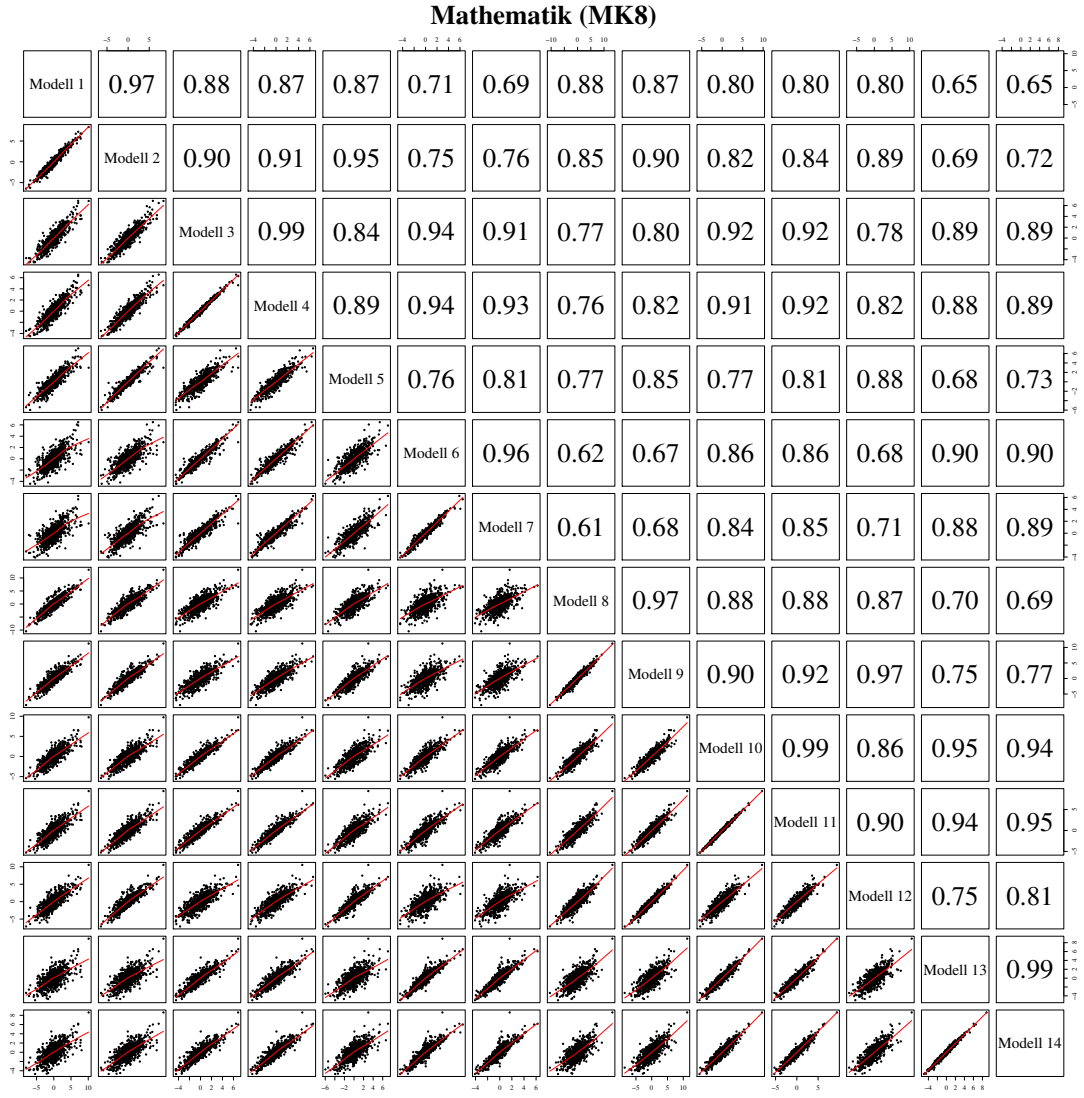

Abbildung 7.13: Korrelationen der adjustierten klassenspezifischen Effektschätzungen zwischen den Modellen im Fachbereich Mathematik (MK8)

(Modell 1) und den Effektschätzungen aus den VAM (Modelle 2 bis 4) unterscheiden sich[12]: Die stärkste Korrelation beträgt $r_{1,2} = .97$, die geringste Korrelation ist $r_{1,4} = .87$.

(2) *VAM vs. CVA*:
Die Korrelationen zwischen den Effektschätzungen der Modelle, die sich hinsichtlich der Berücksichtigung von Kontextvariablen unterscheiden, betragen $r_{2,5} = .95$, $r_{3,6} = .94$ und $r_{4,7} = .93$. Alle drei Korrelationen sind sehr hoch und unterscheiden sich nur gering. Tendenziell zeigt sich eine Abnahme der Korrelation bei steigendem Komplexitätsgrad, d. h. zunehmender Anzahl von Kovariaten in den Modellen.

(3) *Bedingte Unabhängigkeit*:
Auch zwischen den VAM finden sich deutliche Unterschiede hinsichtlich der Korrelationen: Die Effektschätzungen aus Modell 2 und Modell 3 (bzw. Modell 4) korrelieren zu $r_{2,3} = .90$ (bzw. $r_{2,4} = .91$). Demgegenüber besteht zwischen den Effektschätzungen aus den Modellen 3 und 4 mit $r_{3,4} = .99$ eine deutlich stärkere – und zudem die insgesamt stärkste – Korrelation. Folglich zeigt sich eine hohe Stabilität der Effektschätzungen beim Wechsel von Modell 3 zu Modell 4. Mit anderen Worten: Hinsichtlich des Kriteriums Korrelation finden sich keine Hinweise, die die zusätzliche Berücksichtigung des fachspezifischen Vorwissens aus Klassenstufe 3 – zusätzlich zum fachspezifischen Vorwissens aus Klassenstufe 6 *und* aller anderen Kovariaten im Modell (bedingte Unabhängigkeit von MK3) – indizieren. Dies wird weiterhin gestützt durch die lediglich marginalen Unterschiede zwischen den Korrelationen der Modelle 1 und 3 ($r_{1,3} = .88$) sowie der Modelle 1 und 4 ($r_{1,4} = .87$).

Lineare Parametrisierung (ohne Interaktionen). Die Korrelationen zwischen den Effektschätzungen der Modelle mit linearer Parametrisierung (ohne Interaktionen) sind nahezu identisch mit den Korrelationen der Effektschätzungen aus den entsprechenden Modellen mit saturierter und bedingt linearer Parametrisierung (mit Interaktionen). Falls es Unterschiede gibt, so sind diese zumeist sehr gering ($\Delta_r \approx .01$).

[12] Die Korrelation zwischen den klassenspezifischen Effektschätzungen aus einem Modell a und den klassenspezifischen Effektschätzungen aus einem Modell b werde ich nachfolgend mit $r_{a,b}$ abkürzen.

(1) *CAM vs. VAM*:
Die Korrelationen der adjustierten klassenspezifischen Effektschätzungen aus dem CAM und den drei VAM betragen $r_{8,9} = .97$ und $r_{8,10} = r_{8,11} = .88$.

(2) *VAM vs. CVA*:
Die korrelativen Zusammenhänge zwischen den VAM und den jeweils entsprechenden CAM sind auch bei der linearen Parametrisierung sehr hoch und unterscheiden sich nur geringfügig. Die Korrelationen betragen $r_{9,12} = .97$ und $r_{10,13} = r_{11,14} = .95$.

(3) *Bedingte Unabhängigkeit*:
Auch bei der linearen Parametrisierung finden sich deutliche Unterschiede hinsichtlich der Korrelationen zwischen den VAM: Die Effektschätzungen aus den Modellen 9 und 10 (bzw. 11) korrelieren zu $r_{9,10} = .90$ (bzw. $r_{9,11} = .92$). Demgegenüber besteht zwischen den Effektschätzungen aus den Modellen 10 und 11, die sich ausschließlich hinsichtlich der Hinzunahme des klassenspezifischen Vorwissens aus Klassenstufe 3 unterscheiden, mit $r_{10,11} = .99$ wiederum eine deutlich stärkere Korrelation. Dies spricht erneut für die Annahme der bedingten Unabhängigkeit von MK3 gegeben MK6 *und* aller anderen Variablen im Modell.

Bedingt lineare vs. lineare Parametrisierung. Die korrelativen Zusammenhänge der Effektschätzungen aus einander entsprechenden Modellen mit unterschiedlicher Parametrisierung – d. h. Modelle mit dem gleichen Kovariatenset, die sich lediglich hinsichtlich der Parametrisierung (bedingt linear vs. linear) unterscheiden – finden sich in der Diagonale des dritten Bereichs der Korrelationsmatrix. Diese Korrelationen liegen zwischen $r_{\text{min}} = .88$ und $r_{\text{max}} = .92$. Die Korrelationen steigen durch Hinzunahme des fachspezifischen Vorwissens von $r_{1,8} = .88$ auf $r_{3,10} = r_{4,11} = .92$ an. Die Korrelationen sinken wiederum leicht ab, wenn zusätzlich Kontextvariablen in den Modellen berücksichtigt werden: Werden die Kontextvariablen basierend auf dem fachspezifischen Vorwissen aus Klassenstufe 3 einbezogen, beträgt die Korrelation $r_{5,12} = .88$. Bei Hinzunahme von Kontextvariablen basierend auf dem fachspezifischen Vorwissen aus Klassenstufe 6 beträgt die Korrelation $r_{6,13} = .90$. Berücksichtigt man diese Information aus den Klassenstufen 3 *und* 6, resultiert eine Korrelation von $r_{7,14} = .89$.

Korrelationen im Fach Deutsch

Abbildung 7.14 zeigt die Korrelationen der Effektschätzungen einzelner Klassen im Fach Deutsch. Wie bereits im Fachbereich Mathematik finden sich überwiegend hohe Korrelationen, die jedoch auch hier deutliche Unterschiede aufweisen.

Bedingt lineare Parametrisierung (mit Interaktionen). Bei den Modellen mit saturierter und bedingt linearer Parametrisierung (mit Interaktionen) zeigen sich folgende Korrelationen zwischen den resultierenden adjustierten Effektschätzungen:

(1) *CAM vs. VAM*:
Die Korrelationen zwischen den Effektschätzungen des CAM mit denen der drei VAM unterscheiden sich auch im Fach Deutsch. Diese betragen $r_{1,2} = .95$ sowie $r_{1,3} = r_{1,4} = .84$.

(2) *VAM vs. CVA*:
Auch im Fach Deutsch sind die Korrelationen der Effektschätzungen der Modelle, die sich hinsichtlich der Berücksichtigung von Kontextvariablen unterscheiden, sehr hoch. Zwischen den drei Korrelationen $r_{2,5} = .93$ und $r_{3,6} = r_{4,7} = .91$ bestehen lediglich geringe Unterschiede. Dabei zeigt sich wiederum tendenziell eine Abnahme der Korrelation bei steigendem Komplexitätsgrad, d. h. zunehmender Anzahl von Kovariaten in den Modellen.

(3) *Bedingte Unabhängigkeit*:
Des Weiteren finden sich auch hier deutliche Unterschiede zwischen den Korrelationen der Effektschätzungen der VAM: Während die Effektschätzungen aus Modell 2 und 3 (bzw. 4) zu $r_{2,3} = .88$ bzw. $r_{2,4} = .91$ korrelieren, besteht zwischen den Effektschätzungen aus den Modellen 3 und 4 mit $r_{3,4} = .98$ wiederum eine deutlich stärkere Korrelation.

Lineare Parametrisierung (ohne Interaktionen). Auch hier ist das Korrelationsmuster ähnlich in beiden Fachbereichen. Anders als im Fach Mathematik zeigen sich im Fach Deutsch hingegen deutlichere Unterschiede zwischen den Korrelationen unter linearer Parametrisierung (ohne Interaktionen) der Modelle im Vergleich zu den Korrelationen unter bedingt linearer Parametrisierung (mit Interaktionen).

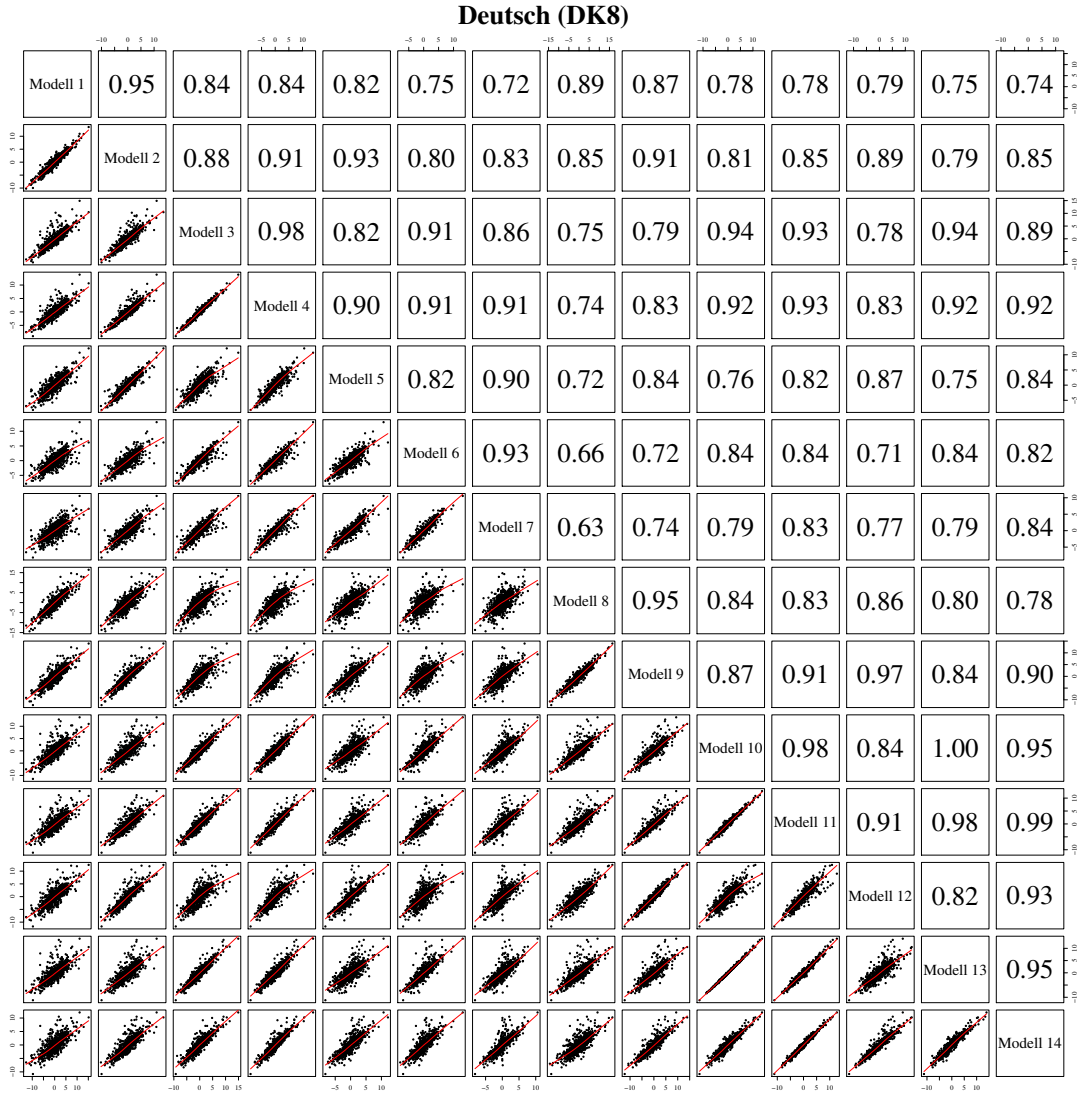

Abbildung 7.14: Korrelationen der adjustierten klassenspezifischen Effektschätzungen zwischen den Modellen im Fachbereich Deutsch (DK8)

(1) *CAM vs. VAM*:
Die Korrelationen der klassenspezifischen Effektschätzungen aus dem CAM und den drei VAM unterscheiden sich auch hier substanziell. Diese betragen jeweils $r_{8,9} = .95$, $r_{8,10} = .84$ und $r_{8,11} = .83$.

(2) *VAM vs. CVA*:
Die Korrelationen zwischen den VAM und den jeweils entsprechenden CAM betragen $r_{9,12} = .97$, $r_{10,13} \approx 1.00$[13] und $r_{11,14} = .99$. Diese korrelativen Zusammenhänge sind somit auch bei der linearen Parametrisierung sehr hoch und unterscheiden sich wiederum nur geringfügig.

(3) *Bedingte Unabhängigkeit*:
Beim Vergleich der Korrelationen zwischen den VAM finden sich auch bei der linearen Parametrisierung deutliche Unterschiede: Die Effektschätzungen aus den Modellen 9 und 10 (bzw. 11) korrelieren zu $r_{9,10} = .87$ (bzw. $r_{9,11} = .91$). Demgegenüber besteht zwischen den Effektschätzungen aus den Modellen 10 und 11, die sich ausschließlich hinsichtlich der Hinzunahme des klassenspezifischen Vorwissens aus Klassenstufe 3 unterscheiden, mit $r_{10,11} = .98$ wiederum eine deutlich stärkere Korrelation.

Bedingt lineare vs. lineare Parametrisierung. Um die Korrelationen der Effektschätzungen aus einander hinsichtlich der Kovariatenselektion entsprechenden Modellen mit unterschiedlicher Parametrisierung beurteilen zu können, betrachten wir auch in Abbildung 7.14 die Diagonale des dritten Bereichs der Korrelationsmatrix. Die Spannweite dieser Korrelation ist im Vergleich zum Fach Mathematik deutlich größer und liegt zwischen $r_{\text{min}} = .84$ und $r_{\text{max}} = .94$. Die Korrelationen steigen durch die Hinzunahme des fachspezifischen Vorwissens von $r_{1,8} = .89$ auf $r_{4,11} = .93$. Wie auch im Fach Mathematik sinken die Korrelationen hingegen wieder ab, wenn zusätzlich Kontextvariablen in den Modellen berücksichtigt werden. Diese Korrelationen zwischen den CVA betragen $r_{5,12} = .87$ und $r_{6,13} = r_{7,14} = .84$.

[13] Der Wert der Korrelation $r_{10,13} = 1.00$ in Abbildung 7.14 kommt durch Rundung der Korrelation zustande. Die Korrelation ist nicht exakt eins, sondern beträgt $r_{10,13} = 0.9980709$.

Zusammenfassung: Korrelationen

Insgesamt sind die Korrelationen zwischen den Effektschätzungen jeweils zweier Modelle zwar hoch. Diese sind jedoch alle kleiner dem Wert 1 ($r \neq 1$) und unterscheiden sich in Abhängigkeit der Kovariaten- und Modellselektion, was beides auf die Sensitivität der Effetkschätzungen hinweist. Für beide Fachbereiche zeigt sich jeweils folgendes Korrelationsmuster der adjustierten klassenspezifischen Effektschätzungen:

Innerhalb einer Parametrisierungsform zeigt sich auch hinsichtlich des Kriteriums der Korrelationen die Sensitivität der adjustierten klassenspezifischen Effektschätzungen gegenüber Modifikationen der Kovariatenselektion. Dabei sind die Korrelationen zwischen CAM und VAM jeweils geringer als zwischen den entsprechenden VAM und CVA. Somit sind die Effektschätzungen sensitiver gegenüber der zusätzlichen Berücksichtigung des fachspezifischen Vorwissens (Hypothese 1.1) als gegenüber der zusätzlichen Berücksichtigung der leistungsmäßigen Klassenkomposition (Hypothese 1.2). Der Vergleich zwischen den beiden Parametrisierungsformen zeigt zudem: Die Korrelationen zwischen Modellen mit jeweils gleichem Kovariatenset, die sich lediglich durch die Modellselektion unterschieden, weisen auf die Sensitivität der adjustierten klassenspezifischen Effektschätzungen gegenüber Modifikationen der Modellselektion hin (Hypothese 2). Des Weiteren zeigen sich auch hier Hinweise auf eine Interaktion zwischen Kovariaten- und Modellselektion (Hypothese 3), allerdings nicht in der angenommenen Richtung: Gemäß Hypothese 3 sollten die Korrelationen zwischen einander hinsichtlich der Kovariatenselektion entsprechenden Modellen stärker werden, je mehr Kovariaten in das Modell hinzugefügt werden. So nehmen die Korrelationen zwischen den VAM im Vergleich zu denen zwischen den CAM zwar zu. Jedoch nehmen diese zwischen den CVA wieder ab. Dieses Ergebnismuster zeigt sich – wie in Hypothese 4 angenommen – gleichermaßen im Fachbereich Mathematik und Deutsch.

7.3.4 Change-Plots

Bisher wurden der Determinationskoeffizient $R^2_{Y|\mathbf{Z}}$ und die Korrelation der adjustierten Effektschätzungen als Kriterien verwendet, um die Ergebnisse der verschiedenen Modelle zu vergleichen. Beide Maße sagen jedoch u. U. wenig darüber aus, welche Konsequenzen die Wahl des Adjustierungsmodells für die individuelle Klasse hat. Die Korrelation der klassenspezifischen Effektschätzungen aus zwei Modellen kann bspw.

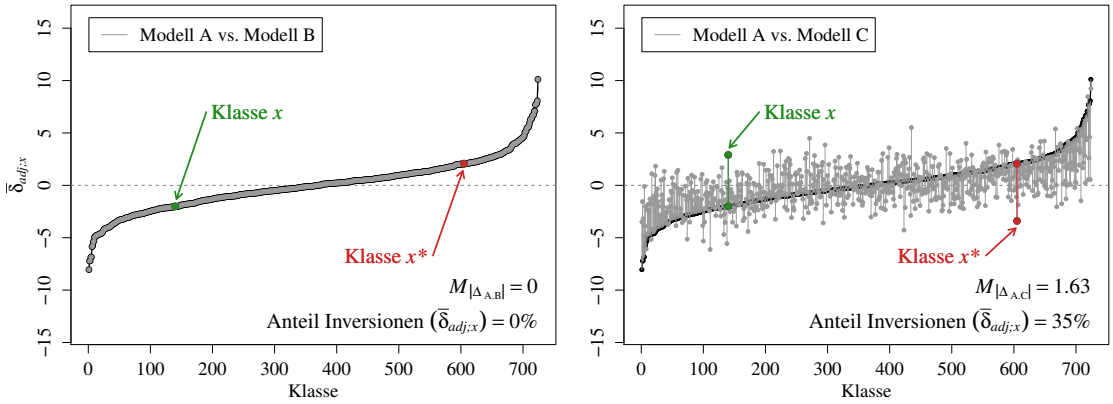

Abbildung 7.15: Zwei Beispiele für Change-Plots. Links: Vergleich von zwei Modellen (A vs. B), deren Effektschätzungen identisch sind. Rechts: Vergleich zweier Modelle (A vs. C) mit starken Unterschieden hinsichtlich der resultierenden Effektschätzungen.

nahezu perfekt positiv sein ($r = 0.99$) und dennoch kann der Wechsel des Analysemodells zu bedeutsamen Unterschieden in den Effektschätzungen einer einzelnen Klasse führen: So kann der Modellwechsel auch bei einer derart hohen Korrelation der Effektschätzungen dennoch für einzelne Klassen zu einer Umkehrung eines positiven in einen negativen Effekt (oder *vice versa*) führen. Die Rückmeldungen der Ergebnisse an die individuellen Klassen sind es jedoch, die im Fokus von landesweiten Vergleichsarbeiten stehen – denn die Klasse ist die primäre Analyseebene im Kontext von Vergleichsarbeiten (vgl. Kapitel 2). Dieser Tatsache soll nachfolgend Rechnung getragen werden, indem die adjustierten Effektschätzungen *einzelner Klassen* aus verschiedenen Modellen miteinander verglichen werden. Zu diesem Zweck verwende ich eine modifizierte Form der Caterpillar-Plots (vgl. Abschnitt 7.3.1), die ich nachfolgend als *Change-Plots* bezeichne. Diese bilden die adjustierten klassenspezifischen Effektschätzungen aus jeweils zwei Modellen sowie (potenzielle) Unterschiede zwischen diesen in einer Grafik ab. Auf diese Weise veranschaulichen Change-Plots die Veränderungen der Effektschätzungen einzelner Klassen, die aus dem Wechsel des Analysemodells resultieren.

Change-Plots: Ein Beispiel. Abbildung 7.15 zeigt exemplarisch zwei Change-Plots, die aus insgesamt drei Modellen – Modelle A, B und C – resultieren. In diesem Beispiel werden die Effektschätzungen von $N = 724$ Klassen verglichen. Wie bei den Caterpillar-Plots sind auch bei dieser Darstellungsform die klassenspezifischen Effekte

$\bar{\delta}_{adj;x}$ auf der Ordinatenachse abgetragen. Die Abszisse repräsentiert wiederum die einzelnen Klassen, die hinsichtlich der Größe der klassenspezifischen Effektschätzungen aus einem Referenzmodell (schwarze Punkte) in eine Rangreihe[14] gebracht wurden. In unserem Beispiel in Abbildung 7.15 ist das Referenzmodell in beiden Fällen jeweils Modell A. Die Standardfehler der klassenspezifischen Effektschätzungen werden *nicht* abgetragen. Stattdessen werden die Effektschätzungen (graue Punkte) eingezeichnet, die aus einem zweiten Modell (Modell B bzw. C) resultieren, wobei die Rangreihe der Klassen unverändert bleibt. Schließlich werden potenzielle Unterschiede zwischen den Effektschätzungen einer Klasse mit einer vertikalen, grauen Linie verbunden. Um das Ausmaß dieser Unterschiede bzw. Veränderungen infolge des Modellwechsels zu quantifizieren, wird außerdem in jedem Change-Plot der Mittelwert des Betrages der Differenzen zwischen den Effektschätzungen angegeben. Zusätzlich wird der Anteil der Inversionen bezüglich der Effektschätzungen $\bar{\delta}_{adj;x}$ angegeben – als Maß für die Richtung der Veränderung der adjustierten Effektschätzungen. Dies ist der prozentuale Anteil der Klassen, bei denen aus dem Referenzmodell eine überdurchschnittliche (positive), jedoch aus dem Vergleichsmodell eine unterdurchschnittliche (negative) Effektschätzung – *und* vice versa – resultiert.

Auf der linken Seite von Abbildung 7.15 werden die Effektschätzungen aus Modell A und Modell B verglichen. Die schwarzen und grauen Punkte überlagern sich perfekt, d. h. die klassenspezifischen Effektschätzungen aus beiden Modellen stimmen exakt überein. Da sich die Effektschätzungen aus beiden Modellen nicht unterscheiden, ist der Mittelwert des Betrages der Differenzen zwischen den Effektschätzungen $M_{|\Delta_{A.B}|} = 0$. Der Change-Plot veranschaulicht weiterhin die Klassen, die unterdurchschnittlich abschneiden ($\bar{\delta}_{adj;x} < 0$). Ein Beispiel dafür ist Klasse x, deren Effektschätzung ($\bar{\delta}_{adj;x} \approx -2$) durch einen grünen Punkt markiert ist. Ebenso werden die Klassen repräsentiert, deren Effektschätzung über dem Durchschnitt liegen ($\bar{\delta}_{adj;x} > 0$). Beispielhaft sei hier die Klasse x^* erwähnt, deren Effektschätzung ($\bar{\delta}_{adj;x^*} \approx 2$) mit roter Farbe gekennzeichnet ist. Jedoch gibt es in diesem Beispiel *keine* Klasse, die aus Modell A eine überdurchschnittliche und aus Modell B eine unterdurchschnittliche Effektschätzung – bzw. umgekehrt – erhält. Demnach liegt der Anteil der Inversionen bezüglich

[14] Im Fokus der nachfolgenden Analyse stehen *Veränderungen* der adjustierten Effektschätzungen einzelner Klassen infolge der Modifikation des Adjustierungsmodells. Dabei ist die Rangreihung der Klassen per se vernachlässigbar. Die Klassen werden ausschließlich aus Gründen der besseren Darstellung dieser Veränderungen in eine Rangreihe gebracht.

der Effektschätzungen $\bar{\delta}_{adj;x}$ bei 0%.

Die rechte Seite in Abbildung 7.15 zeigt das Ergebnis des Modellvergleichs von Modell A vs. Modell C. Da wiederum Modell A das Referenzmodell ist, sind die schwarzen Punkte identisch mit denen der linken Grafik. Die grauen Punkte repräsentieren nun jedoch die klassenspezifischen Effektschätzungen aus Modell C. Diese sind zusätzlich durch graue Linien mit den entsprechenden Effektschätzungen der jeweils gleichen Klasse aus Modell A verbunden, falls diese voneinander abweichen. Je weiter die Effektschätzungen aus beiden Modellen pro Klasse voneinander abweichen, desto länger sind die grauen Linien und desto „unruhiger" wird die Grafik. Der Mittelwert des Betrages dieser Abweichungen beträgt im vorliegenden Beispiel $M_{|\Delta_{A,C}|} = 1.63$. Auch kehren sich die Effektschätzungen unserer zwei Beispielklassen x und x^* bezüglich des Vorzeichens um: Klasse x, deren Effektschätzung aus Modell A unterdurchschnittlich ausfiel, erhält nun eine positive Effektschätzung ($\bar{\delta}_{adj;x} \approx 3$). Bei Klasse x^* hingegen findet sich nun – anstatt des ursprünglichen positiven Effekts – eine negative Effektschätzung ($\bar{\delta}_{adj;x^*} \approx -3$). Insgesamt liegen bei 35% der betrachteten $N = 724$ Klassen Inversionen bezüglich der Richtung der Effektschätzung beim Wechsel von Modell A zu Modell C vor. Demnach liegt bei mehr als $n = 250$ Klassen eine Inversion der Effektschätzung vor. Zusammenfassend zeigt sich in diesem Beispiel somit eine hohe Sensitivität der adjustierten klassenspezifischen Effektschätzungen beim Wechsel des Analysemodells von Modell A zu Modell C.

Change-Plots im Fach Mathematik

Die Abbildungen 7.16 bis 7.22 zeigen die adjustierten klassenspezifischen Effektschätzungen der unterschiedlichen Modelle für insgesamt $N = 724$ Thüringer Klassen der Klassenstufe 8, bei denen die Mathematikleistung mittels des Kompetenztests Mathematik (MK8) erhoben wurde. Unterschiede zwischen den Effektschätzungen, die aus jeweils zwei Modellen resultieren, sind je nach Parametrisierung durch verschiedene Farben gekennzeichnet: (a) Beim Vergleich zwischen Modellen mit bedingt linearer Parametrisierung (mit Interaktionen) wird die Farbe Cyan, (b) beim Vergleich linearer Modelle (ohne Interaktionen) hingegen die Farbe Rot verwendet. Schließlich sind (c) Vergleiche zwischen Modellen unterschiedlicher Parametrisierung (bedingt linear vs. linear) in grau dargestellt. Die Modellvergleiche werden nachfolgend in eben dieser Reihenfolge – (a), (b) und schließlich (c) – dargestellt.

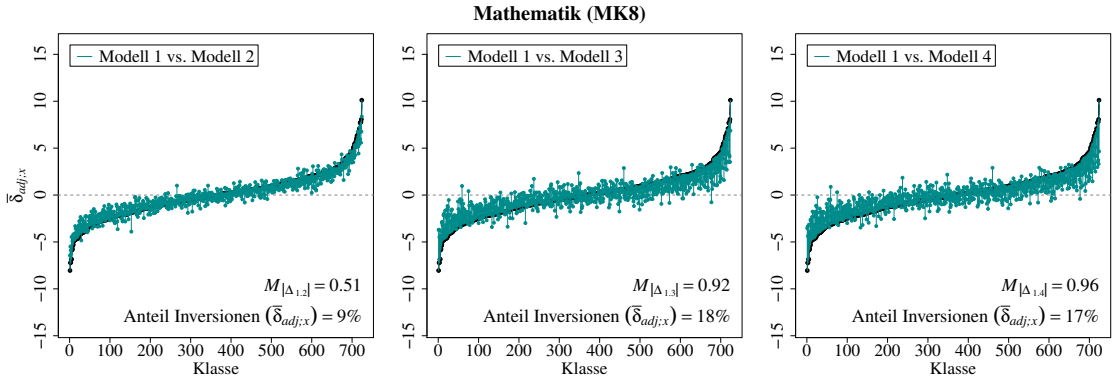

Abbildung 7.16: Change-Plots im Fach Mathematik (MK8): CAM *versus* VAM (saturierte und bedingt lineare Parametrisierung mit Interaktionen)

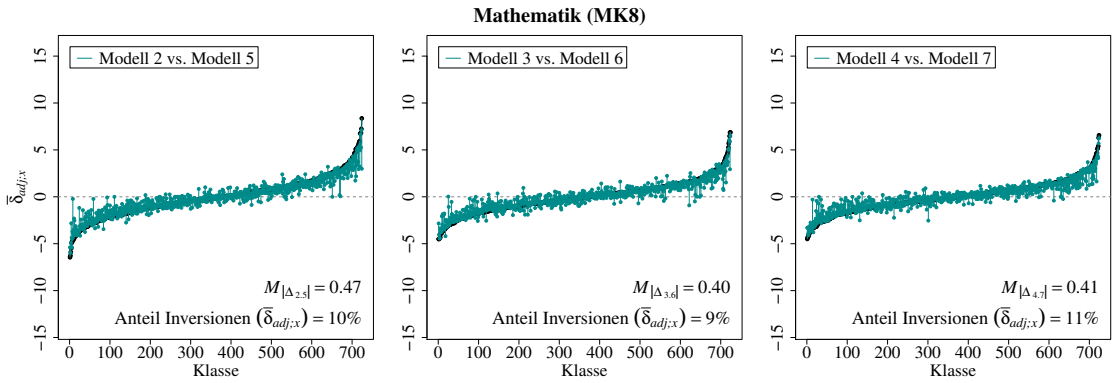

Abbildung 7.17: Change-Plots im Fach Mathematik (MK8): VAM *versus* CVA (bedingt lineare Parametrisierung mit Interaktionen)

Bedingt lineare Parametrisierung (mit Interaktionen). Zunächst betrachten wir die Modelle 1 bis 7 mit saturierter und bedingt linearer Parametrisierung (inklusive Interaktionen). Da nachfolgend Effektschätzungen aus Modellen mit jeweils identischer Parametrisierung verglichen werden, sind potenzielle Unterschiede in den Effektschätzungen nicht auf die Modellselektion, sondern auf die Wahl des Kovariatensets attribuierbar.

(1) *CAM vs. VAM*:
Abbildung 7.16 zeigt die drei Change-Plots für den Wechsel von Modell 1 (CAM) zu je einem Modell des Typs VAM, bei dem zusätzlich das fachspezifische Vorwissen berücksichtigt wird. Die Referenz ist somit bei jedem der drei Vergleiche das erste Modell.

Wird zusätzlich zu den Kovariaten im Modell 1 auch das fachspezifische Vorwissen aus Klassenstufe 3 (MK3) in das Modell aufgenommen (Modell 2), so zeigen sich augenfällig Veränderungen in den adjustierten klassenspezifischen Effektschätzungen. Der Mittelwert des Betrages dieser Abweichungen ist $M_{|\Delta_{1.2}|} = 0.51$. Der Anteil der Inversionen hinsichtlich der Effektschätzungen $\bar{\delta}_{adj;x}$ beträgt 9% ($n = 62$ Klassen). Wird hingegen MK6 anstatt MK3 zusätzlich in dem Adjustierungsmodell berücksichtigt (Modell 1 vs. Modell 3), so beträgt die durchschnittliche Abweichung $M_{|\Delta_{1.3}|} = 0.92$ und der Anteil der Inversionen in $\bar{\delta}_{adj;x}$ verdoppelt sich auf 18% (bzw. $n = 131$ der insgesamt $N = 724$ betrachteten Klassen). Ein dazu sehr ähnliches Bild ergibt sich, wenn beide Vorwissensvariablen – sowohl MK3 als auch MK6 – zusätzlich in das Modell aufgenommen werden (Modell 1 vs. Modell 4). Hier resultiert eine durchschnittliche Abweichung von $M_{|\Delta_{1.4}|} = 0.96$ und der Anteil der Inversionen beträgt 17% ($n = 124$ Klassen).

(2) *VAM vs. CVA*:
In Abbildung 7.17 sind die Change-Plots dargestellt, die beim Wechsel vom VAM zum CVA resultieren. Hier wird die Frage adressiert, welchen Einfluss die zusätzliche Berücksichtigung von Klassenkompositionsmerkmalen auf die Veränderungen der adjustierten Effektschätzungen haben. Dabei handelt es sich wiederum um einen paarweisen Vergleich jeweils genesteter Modelle (vgl. Abschnitt 7.3.2), d. h. im Einzelnen: Modell 2 vs. 5, Modell 3 vs. 6 und Modell 4 vs. 7.
Werden zusätzlich zu den ursprünglichen Kovariaten *und* zum fachspezifischen Vorwissen aus Klassenstufe 3 (MK3) auch die entsprechenden Klassenkompositionsmerkmale hinsichtlich des Vorwissens aus Klassenstufe 3 in das Modell aufgenommen (Modell 2 vs. Modell 5), so liegt der Mittelwert des Betrages der Differenzen zwischen den Effektschätzungen bei $M_{|\Delta_{2.5}|} = 0.47$. Gleichzeitig wird bei 10% aller Klassen ($n = 76$) ein positiver zu einem negativen Effekt und umgekehrt. Etwas geringer ist die Veränderung infolge der Hinzunahme des fachspezifischen Vorwissens in Klassenstufe 6 und der entsprechenden Kompositionsmerkmale (Modell 3 vs. Modell 6): Hier beträgt die durchschnittliche Differenz der Effektschätzungen $M_{|\Delta_{3.6}|} = 0.40$. Zudem finden sich bei 9% ($n = 68$) der Klassen Inversionen der Effekte beim Wechsel von Modell 3 zu Modell 6. Und auch für den Fall, dass Informationen sowohl zum fachspezifischen Vorwissen aus Klassenstufe 3 (MK3) als auch Klassenstufe 6 (MK6) sowie der entsprechen-

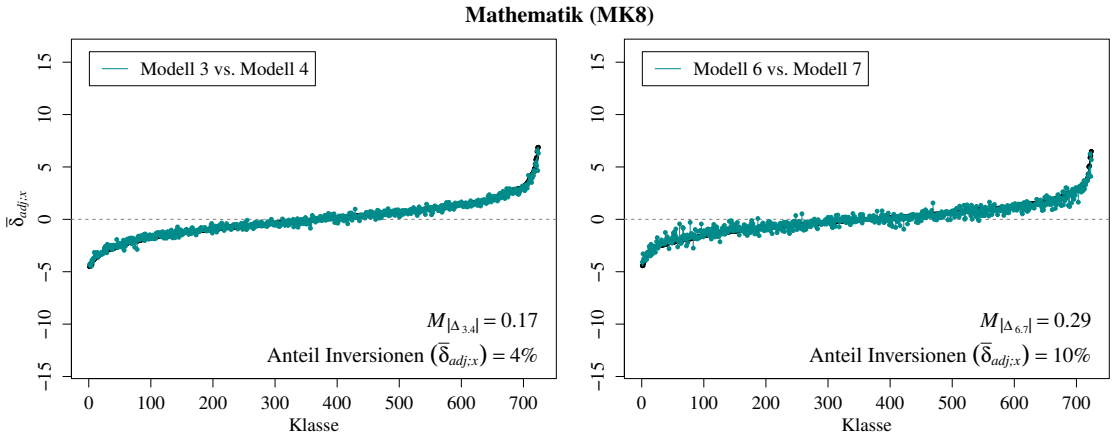

Abbildung 7.18: Change-Plots im Fach Mathematik (MK8): Modelle mit *versus* ohne MK3 (bedingt lineare Parametrisierung mit Interaktionen)

den Klassenkomposition zur Verfügung stehen (Modell 4 vs. Modell 7), liegt die durchschnittliche Differenz der Effektschätzungen bei $M_{|\Delta_{4.7}|} = 0.41$, wobei für etwa $n = 79$ Klassen (11%) eine Inversion der Effekte beobachtbar ist.

Zusammenfassend zeigt sich somit zwischen diesen drei Modellvergleichen ein insgesamt homogeneres Bild als in Abbildung 7.16. Mit anderen Worten: Unabhängig davon, ob die leistungsmäßige Klassenkomposition hinsichtlich MK3, MK6 *oder* sowohl MK3 als auch MK6 zusätzlich im Adjustierungsmodell berücksichtigt wird, zeigt sich eine vergleichbar starke Sensitivität der adjustierten klassenspezifischen Effektschätzungen. Zudem sind die Veränderungen infolge der zusätzlichen Berücksichtigung von Klassenkompositionsmerkmalen jeweils ähnlich bzw. geringer als bei der Hinzunahme des fachspezifischen Vorwissens.

(3) *Bedingte Unabhängigkeit*:

Ist es hinreichend, anstatt beider Vorwissensvariablen (MK3 und MK6) nur das fachspezifische Vorwissen aus Klassenstufe 6 (MK6) zusätzlich in das Modell aufzunehmen? Bei dieser Frage geht es wiederum um die bedingte Unabhängigkeit der Testwertvariablen MK8 von MK3 gegeben MK6 und der restlichen Kovariaten im Adjustierungsmodell. Wäre dies der Fall, so sollte der Wechsel von Modell 3 zu Modell 4 zu keinen Veränderungen in den adjustierten Effektschätzungen führen. Der entsprechende Vergleich der adjustierten Effektschätzungen aus beiden Modellen ist in Abbildung 7.18 (linke Grafik) wiedergegeben. Das Ergebnis spricht für die postulierte bedingte Unabhängigkeit, da der Modellwechsel

mit lediglich kleineren Veränderungen einhergeht: Die durchschnittliche Abweichung liegt hier bei $M_{|\Delta_{3,4}|} = 0.17$ mit einem lediglich 4%-igem Anteil an Inversionen ($n = 27$ Klassen). Hingegen zeigt sich ein anderes Bild, wenn man die Sensitivität der Effektschätzungen bei Hinzunahme der entsprechenden Kompositionsmerkmale betrachtet: Auf der rechten Seite in Abbildung 7.18 werden die Effektschätzungen aus den Modellen 6 und 7 miteinander verglichen. Die durchschnittliche Veränderung der Effektschätzungen beträgt hier $M_{|\Delta_{6,7}|} = 0.29$, wobei sich bei 10% der Klassen $n = 70$ die Richtung des Effekts (positiv vs. negativ) umkehrt.

Lineare Parametrisierung (ohne Interaktionen). Nachfolgend werden die entsprechenden Ergebnisse der Modelle mit linearer Parametrisierung einer vergleichenden Betrachtung unterzogen. Auch hier steht die Sensitivität der adjustierten Effektschätzungen hinsichtlich der Kovariatenselektion im Fokus der Betrachtung.

(1) *CAM vs. VAM*:
In Abbildung 7.19 sind die drei Change-Plots wiedergegeben, die aus dem Wechsel des Adjustierungmodells von Modell 8 (CAM) zu je einem Modell des Typs VAM (Modell 9, 10 bzw. 11) resultieren. Der linke Change-Plot zeigt die Veränderungen der adjustierten klassenspezifischen Effektschätzungen infolge der zusätzlichen Berücksichtigung von MK3 (Modell 8 vs. Modell 9). Der Mittelwert des Betrages dieser Veränderungen ist $M_{|\Delta_{8,9}|} = 0.61$, wobei der Anteil der Inversionen bei 9% ($n = 63$) liegt. Wird hingegen MK6 anstatt MK3 zusätzlich im Adjustierungsmodell berücksichtigt (Modell 8 vs. Modell 10), so beträgt die durchschnittliche Abweichung $M_{|\Delta_{8,10}|} = 1.09$. Zudem verdoppelt sich – wie auch bei den bedingt linear parametrisierten Modellen (vgl. Abbildung 7.16) – der Anteil der Inversionen in $\bar{\delta}_{adj;x}$ auf 17% und betrifft somit $n = 126$ der insgesamt $N = 724$ betrachteten Klassen. Ein sehr ähnliches Bild zeigt sich, wenn sowohl MK3 als auch MK6 zusätzlich in das Modell aufgenommen werden (Modell 8 vs. Modell 11). Dabei resultiert eine durchschnittliche Veränderung von $M_{|\Delta_{8,11}|} = 1.14$ und der Anteil der Inversionen beträgt gleichfalls 17% ($n = 125$ Klassen).

(2) *VAM vs. CVA*:
Welchen Einfluss hat nun die zusätzliche Berücksichtigung der leistungsmäßi-

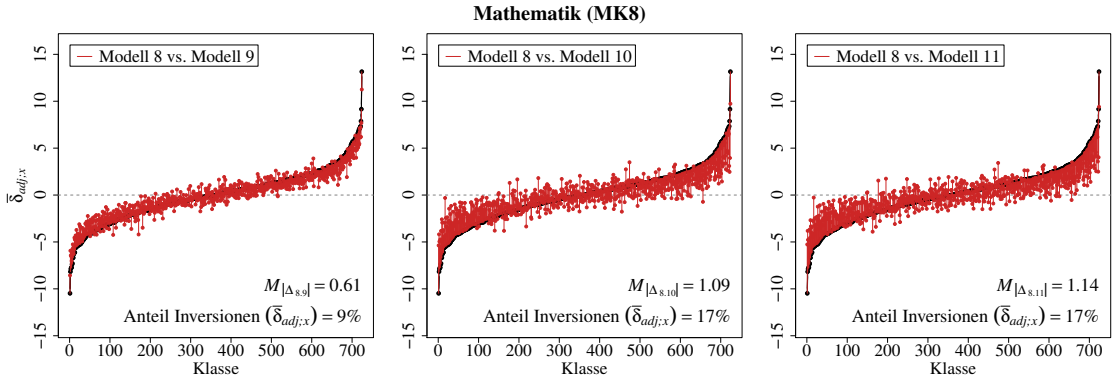

Abbildung 7.19: Change-Plots im Fach Mathematik (MK8): CAM *versus* VAM (lineare Parametrisierung ohne Interaktionen)

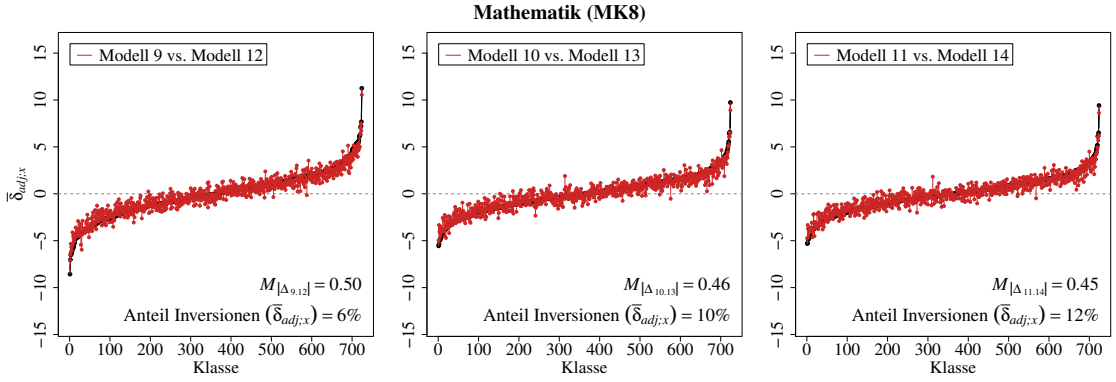

Abbildung 7.20: Change-Plots im Fach Mathematik (MK8): VAM *versus* CVA (lineare Parametrisierung ohne Interaktionen)

gen Klassenkomposition auf die Sensitivität der adjustierten Effektschätzungen in den linearen Adjustierungsmodellen? Die Ergebnisse der diese Frage adressierenden Modellvergleiche sind in Abbildung 7.20 dargestellt. Dabei handelt es sich um den paarweisen Vergleich der folgenden genesteten Modelle: Modell 9 vs. 12, Modell 10 vs. 13 sowie Modell 11 vs. 14.

Werden zusätzlich zu den restlichen Kovariaten *und* zum fachspezifischen Vorwissen aus Klassenstufe 3 (MK3) auch die entsprechenden Klassenkompositionsmerkmale hinsichtlich des Vorwissens aus Klassenstufe 3 in das Modell aufgenommen (Modell 9 vs. Modell 12), so liegt der Mittelwert des Betrages der Differenzen zwischen den Effektschätzungen bei $M_{|\Delta_{9.12}|} = 0.50$. Gleichzeitig wird bei 6% aller Klassen ($n = 45$) ein positiver zu einem negativen Effekt und umgekehrt. Etwas geringer ist die Veränderung infolge der Hinzunahme des

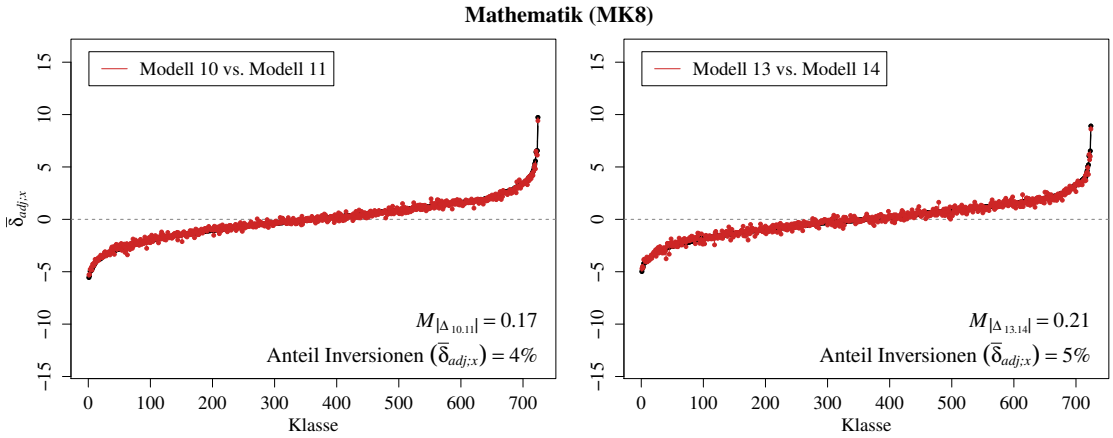

Abbildung 7.21: Change-Plots im Fach Mathematik (MK8): Modelle mit *versus* ohne MK3 (lineare Parametrisierung ohne Interaktionen)

fachspezifischen Vorwissens in Klassenstufe 6 und der entsprechenden Kompositionsmerkmale (Modell 10 vs. Modell 13): Hier beträgt die durchschnittliche Differenz der Effektschätzungen $M_{|\Delta_{10.13}|} = 0.46$. Zudem finden sich bei 10% der Klassen ($n = 76$) Inversionen der Effekte beim Wechsel von Modell 3 zu Modell 6. Und auch für den Fall, dass Informationen sowohl zum fachspezifischen Vorwissen aus Klassenstufe 3 (MK3) als auch Klassenstufe 6 (MK6) sowie der entsprechenden Klassenkomposition zur Verfügung stehen (Modell 11 vs. 14), liegt die durchschnittliche Differenz der Effektschätzungen bei $M_{|\Delta_{11.14}|} = 0.45$, wobei für $n = 84$ Klassen (12%) eine Inversion der Effekte beobachtbar ist.

Summa summarum zeigt sich – wie bereits beim entsprechenden Vergleich der bedingt linearen Adjustierungsmodelle – ein homogeneres Bild als in Abbildung 7.19. Anders formuliert: Unabhängig davon, ob die leistungsmäßige Klassenkomposition hinsichtlich MK3, MK6 *oder* sowohl MK3 als auch MK6 zusätzlich im linearen Adjustierungsmodell berücksichtigt wird, zeigt sich eine vergleichbar starke Sensitivität der adjustierten klassenspezifischen Effektschätzungen. Des Weiteren sind die Veränderungen infolge der zusätzlichen Berücksichtigung von Klassenkompositionsmerkmalen jeweils geringer als bei der Hinzunahme des fachspezifischen Vorwissens.

(3) *Bedingte Unabhängigkeit*:

Hinsichtlich der bedingten Unabhängigkeit der Testwertvariablen MK8 von MK3 gegeben MK6 und der restlichen Kovariaten im linearen Adjustierungsmodell

(ohne Interaktionen) zeigt sich ein ähnliches Ergebnis wie bereits bei der bedingt linearen Parametrisierung (vgl. Abbildung 7.18): Der entsprechende Vergleich der Effektschätzungen aus den Modellen 10 und 11 ist in Abbildung 7.21 (linke Grafik) wiedergegeben. Das Ergebnis spricht für die postulierte bedingte Unabhängigkeit, da der Modellwechsel mit lediglich kleineren Veränderungen einhergeht: Die durchschnittliche Abweichung liegt hier bei $M_{|\Delta_{10.11}|} = 0.17$ mit einem lediglich 4%-igem Anteil an Inversionen ($n = 31$ Klassen). Ein vergleichbares Ergebnis zeigt sich, wenn man die Sensitivität der Effektschätzungen bei Hinzunahme der entsprechenden Kompositionsmerkmale betrachtet: Auf der rechten Seite in Abbildung 7.21 ist der Vergleich der Effektschätzungen aus den Modellen 13 und 14 mit linearer Parametrisierung dargestellt. Die durchschnittliche Veränderung der Effektschätzungen bei $M_{|\Delta_{13.14}|} = 0.21$, wobei sich bei 5% der Klassen ($n = 31$) die Richtung des Effekts (positiv oder negativ) umkehrt.

In Relation zum zuvor dargestellten Vergleich der Modelle mit saturierter und bedingt linearer Parametrisierung (mit Interaktionen) fällt auf, dass sich zwar die durchschnittlichen Beträge der Abweichungen der Effektschätzungen unterscheiden. Diese sind stets größer beim Vergleich linear parametrisierter Modelle als beim Vergleich der entsprechenden Modellen mit dem jeweils gleichem Kovariatenset, aber mit saturierter und bedingt linearer Parametrisierung. Jedoch ist der Anteil der Klassen, bei denen sich die adjustierten Effektschätzungen infolge des Modellwechsels von einem positiven in einen negativen Effekt (und vice versa) umkehrt, jeweils fast identisch. So ist bspw. die durchschnittliche Abweichung der Effekte $M_{|\Delta_{1.3}|} = 0.92$, wenn MK6 zusätzlich zu den anderen Kovariaten in das Modell mit bedingt linearer Parametrisierung aufgenommen wird (vgl. Abbildung 7.16). Die Abweichung zwischen den beiden hinsichtlich der Kovariatenselektion einander entsprechenden Modellen mit linearer Parametrisierung beträgt hingegen $M_{|\Delta_{8.10}|} = 1.09$ (vgl. Abbildung 7.19). Jedoch ist der Anteil der Inversionen mit 18% bzw. 17% ($n = 131$ bzw. $n = 126$ der insgesamt $N = 724$ betrachteten Klassen) nahezu gleich. Dieses Ergebnismuster ist konkordant mit dem Ergebnis, dass die Varianzen bei einander hinsichtlich der Kovariatenselektion entsprechenden Modellen jeweils größer in linear paramerisierten Modellen sind im Vergleich zu den bedingt linearen Modellen (vgl. Abschnitt 7.3.1).

Bedingt lineare vs. lineare Parametrisierung. Bisher wurden solche Modelle verglichen, die sich ausschließlich hinsichtlich der berücksichtigten Kovariaten unterscheiden. Die Parametrisierung der Modelle war jedoch jeweils konstant. Dabei stand die Sensitivität der adjustierten Effektschätzungen infolge der Modifikation der Kovariatenselektion im Fokus der Betrachtung. Nun sollen schließlich auch die verschiedenen Parametrisierungen einer vergleichenden Betrachtung unterzogen werden, um die Sensitivität der adjustierten Effektschätzungen infolge der Modifikation der Modellselektion beurteilen zu können.

Abbildung 7.22 zeigt die Veränderungen der adjustierten klassenspezifischen Effektschätzungen infolge des Wechsels der Parametrisierung. Welchen Einfluss hat die Modellselektion – saturierte vs. lineare Parametrisisierung – auf die adjustierten klassenspezifischen Effektschätzungen beim CAM (Modell 1 vs. Modell 8)? Die durchschnittliche Veränderung der Effektschätzungen beträgt hier $M_{|\Delta_{1.8}|} = 1.05$ und für 18% der Klassen ($n = 128$) kehrt sich die Richtung des Effekts um. Beim Vergleich der zwei VAM, die beide das fachspezifische Vorwissen der Jahrgangsstufen 3 und 6 zusätzlich zu den Kovariaten aus dem CAM enthalten (Modell 4 vs. Modell 11), ist diese Veränderung hingegen geringer: Der Mittelwert des Betrages dieser Veränderungen liegt bei $M_{|\Delta_{4.11}|} = 0.56$. Auch der Anteil der Inversionen ist hier mit 13% bzw. $n = 91$ Klassen geringer als beim vorangegangen Modellvergleich. Schließlich zeigt ein Vergleich der zwei CVA, die sich hinsichtlich der Parametrisierung unterscheiden (Modell 7 vs. Modell 14): Die durchschnittliche Veränderung ist $M_{|\Delta_{7.14}|} = 0.55$. Der Anteil der Inversionen beträgt 15%, d. h. $n = 108$ Klassen.

Insgesamt ist der Einfluss der Parametrisierung am stärksten für die CAM (d. h. Modelle ohne das fachspezifische Vorwissen). Dieser Einfluss wird geringer, wenn das fachspezifische Vorwissen im Modell enthalten ist (VAM) bzw. auch die leistungsmäßige Klassenkomposition berücksichtigt wird (CVA). Vergleicht man hingegen VAM und CVA, so ist – bis auf eine Ausnahme (VAM: Modell 2 vs. 9 und CVA: Modell 5 vs. 12) – eine höhere Stabilität der VAM beobachtbar: So führt der Parametrisierungswechsel von Modell 4 zu Modell 11 (VAM) zwar zu ähnlich großen Unterschieden der Effektschätzungen. Jedoch resultiert ein kleinerer Anteil an Inversionen als beim entsprechenden Wechsel von Modell 7 zu Modell 14 (CVA). Diese tendenziell größere Stabilität der VAM gegenüber des Wechsels der Parametrisierung hinsichtlich der Veränderung der Effektschätzungen einzelner Klassen ist somit vergleichbar mit dem Ergebnismuster hinsichtlich der Korrelationen (vgl. Abschnitt 7.3.3).

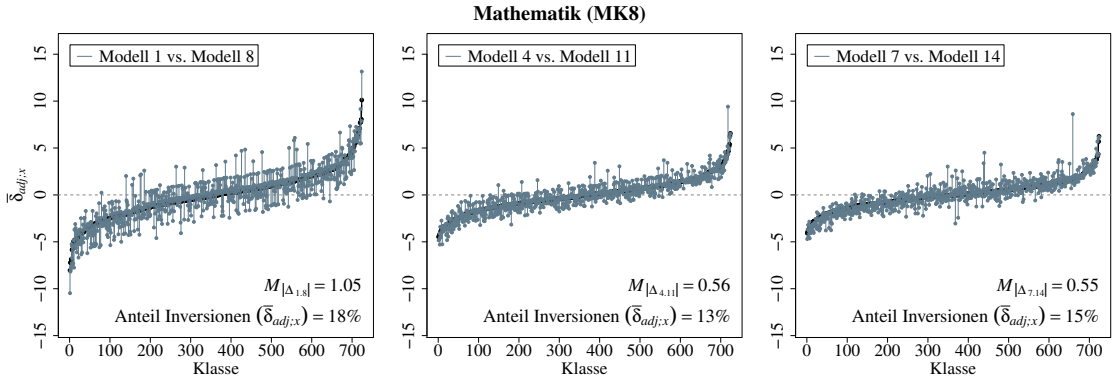

Abbildung 7.22: Change-Plots im Fach Mathematik (MK8): Bedingt lineare *versus* lineare Parametrisierung

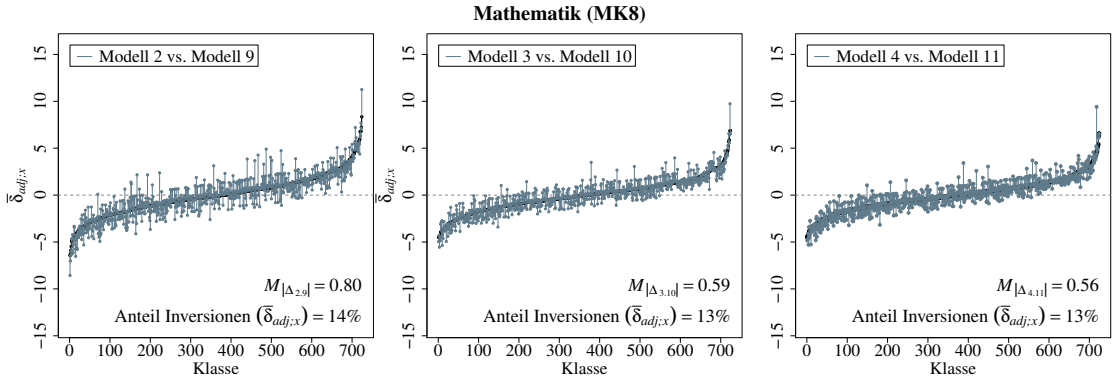

Abbildung 7.23: Change-Plots im Fach Mathematik (MK8): VAM mit bedingt linearer *versus* linearer Parametrisierung

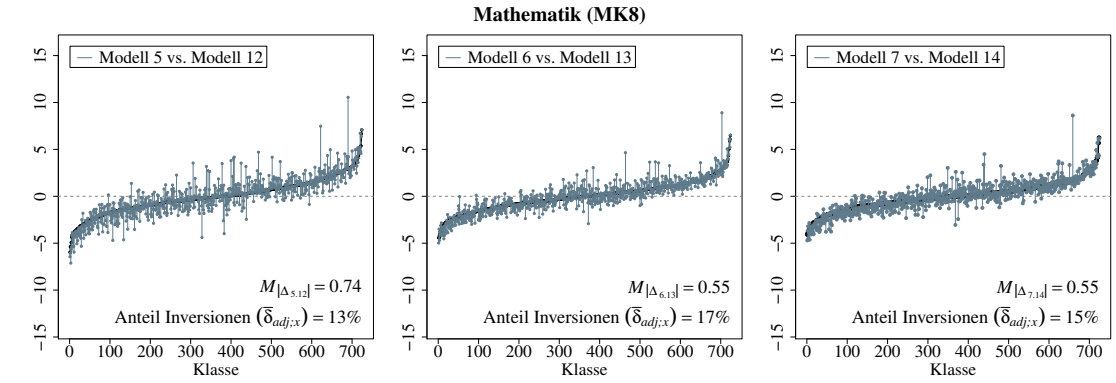

Abbildung 7.24: Change-Plots im Fach Mathematik (MK8): CVA mit bedingt linearer *versus* linearer Parametrisierung

Zudem zeigt sich jedoch gleichfalls eine deutliche Variabilität – in Bezug auf die Sensitivität der Effektschätzungen – *zwischen* den einzelnen VAM respektive CVA. Die zugehörige grafische Darstellung dieser Modellvergleiche findet sich in den Abbildungen 7.23 und 7.24. Innerhalb der Modellklasse VAM schwanken die durchschnittlichen Abweichungen zwischen maximal $M_{|\Delta_{2.9}|} = 0.80$ und minimal $M_{|\Delta_{4.11}|} = 0.56$. Der Anteil der Inversionen beträgt zwischen 14% ($n = 101$) beim Wechsel von Modell 2 zu 9 und 13% ($n = 91$) beim Wechsel von Modell 4 zu 11 der Klassen. Innerhalb der Modellklasse CVA variieren die durchschnittlichen Abweichungen zwischen $M_{|\Delta_{5.12}|} = 0.74$ und $M_{|\Delta_{6.13}|} = 0.55$, wobei gleichzeitig zwischen 13% ($n = 94$) und 17% ($n = 121$) der Klassen von einer Inversion der Effektschätzung betroffen sind.

Change-Plots im Fach Deutsch

Die Abbildungen 7.25 bis 7.31 zeigen die aus den verschiedenen Modellen resultierenden adjustierten Effektschätzungen für jeweils $N = 702$ Thüringer Klassen der Klassenstufe 8, bei denen die Deutschleistung mittels des Kompetenztests Deutsch (DK8) erhoben wurde. Wie bereits für Mathematik verwende ich auch in der nachfolgenden Ergebnisdarstellung für den Fachbereich Deutsch je nach Parametrisierung unterschiedliche Farben: (a) Beim Vergleich zwischen Modellen mit bedingt linearer Parametrisierung (mit Interaktionen) wird die Farbe Cyan, (b) beim Vergleich linearer Modelle (ohne Interaktionen) hingegen die Farbe Rot verwendet. Schließlich sind (c) Vergleiche zwischen Modellen unterschiedlicher Parametrisierung (bedingt linear vs. linear) in grau dargestellt. Die Modellvergleiche werden nachfolgend in eben dieser Reihenfolge – (a), (b) und schließlich (c) – dargestellt.

Bedingt lineare Parametrisierung (mit Interaktionen). Nachfolgend werden die Ergebnisse des Vergleichs der Modelle 1 bis 7 (saturierte und bedingt lineare Parametrisierung mit Interaktionen) dargelegt.

(1) *CAM vs. VAM*:
Abbildung 7.25 zeigt die Change-Plots für den Wechsel von Modell 1 (CAM) zu je einem Modell des Typs VAM, bei dem zusätzlich das fachspezifische Vorwissen berücksichtigt wird. Folglich bildet Modell 1 jeweils die Referenz im Rahmen dieser drei Modellvergleiche.

Wird zusätzlich zu den Kovariaten im CAM (Modell 1) auch das fachspezifische Vorwissen aus Klassenstufe 3 (DK3) in das Modell aufgenommen (Modell 2), so zeigen sich deutliche Veränderungen in den adjustierten klassenspezifischen Effektschätzungen. Der Mittelwert des Betrages dieser Abweichungen ist $M_{|\Delta_{1.2}|} = 0.91$. Außerdem sind 9% bzw. $n = 65$ von insgesamt $N = 702$ Thüringer Klassen von Inversionen der Effektschätzungen betroffen. Wird DK6 (anstatt DK3) zusätzlich im Adjustierungsmodell berücksichtigt (Modell 1 vs. Modell 3), so beträgt die durchschnittliche Abweichung sogar $M_{|\Delta_{1.3}|} = 1.43$. Zudem nimmt hier auch der Anteil der Klassen mit Inversionen deutlich zu (17% bzw. $n = 117$ Klassen). Ein sehr ähnliches Ergebnismuster ergibt sich, wenn beide Vorwissensvariablen – sowohl DK3 als auch DK6 – zusätzlich in das Modell aufgenommen werden (Modell 1 vs. Modell 4). Hier resultiert eine durchschnittliche Abweichung von $M_{|\Delta_{1.4}|} = 1.54$, wobei der Anteil der Inversionen 19% ($n = 130$ Klassen) beträgt.

(2) *VAM vs. CVA*:
In Abbildung 7.26 sind die Change-Plots dargestellt, die beim Wechsel vom VAM zum CVA resultieren. Mittels dieses Vergleichs wird wiederum die Frage adressiert, welchen Einfluss die zusätzliche Berücksichtigung von Klassenkompositionsmerkmalen auf die Veränderungen der adjustierten Effektschätzungen haben. Auch hier handelt es sich um einen paarweisen Vergleich jeweils genesteter Modelle (vgl. Abschnitt 7.3.2). Im Einzelnen handelt es sich um die folgenden Modellvergleiche: Modell 2 vs. 5, Modell 3 vs. 6 und Modell 4 vs. 7.
Hier zeigt sich – ähnlich zu dem Ergebnismuster im Fach Mathematik – zwischen den drei Modellvergleichen ein insgesamt homogeneres Bild als beim Wechsel vom CAM zum VAM (Abbildung 7.25). Zudem sind die Veränderungen infolge der zusätzlichen Berücksichtigung von Klassenkompositionsmerkmalen jeweils geringer als beim Wechsel vom CAM zum VAM: Werden zusätzlich zu den ursprünglichen Kovariaten *und* zum fachspezifischen Vorwissen aus Klassenstufe 3 (DK3) auch die entsprechenden Klassenkompositionsmerkmale hinsichtlich des Vorwissens aus Klassenstufe 3 in das Modell aufgenommen (Modell 2 vs. Modell 5), so liegt der Mittelwert des Betrages der Differenzen zwischen den Effektschätzungen bei $M_{|\Delta_{2.5}|} = 0.77$. Gleichzeitig wird bei 9% aller Klassen ($n = 62$) ein positiver zu einem negativen Effekt und umgekehrt. Etwas stär-

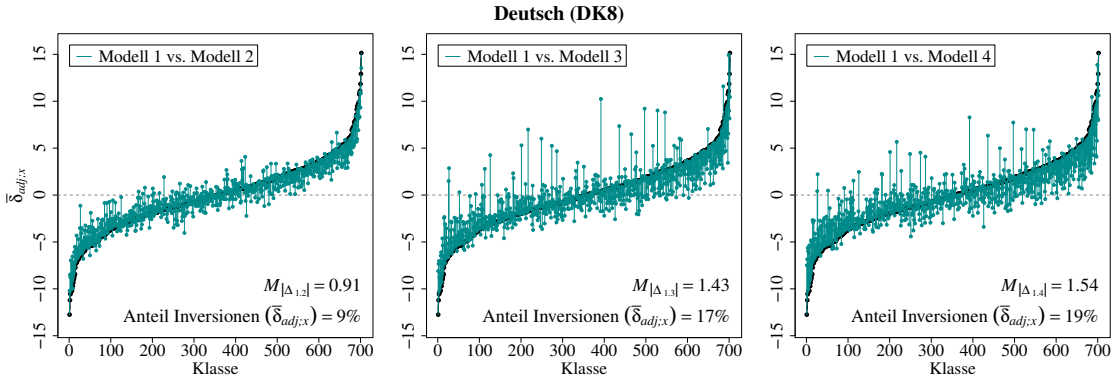

Abbildung 7.25: Change-Plots im Fach Deutsch (DK8): CAM *versus* VAM (saturierte und bedingt lineare Parametrisierung mit Interaktionen)

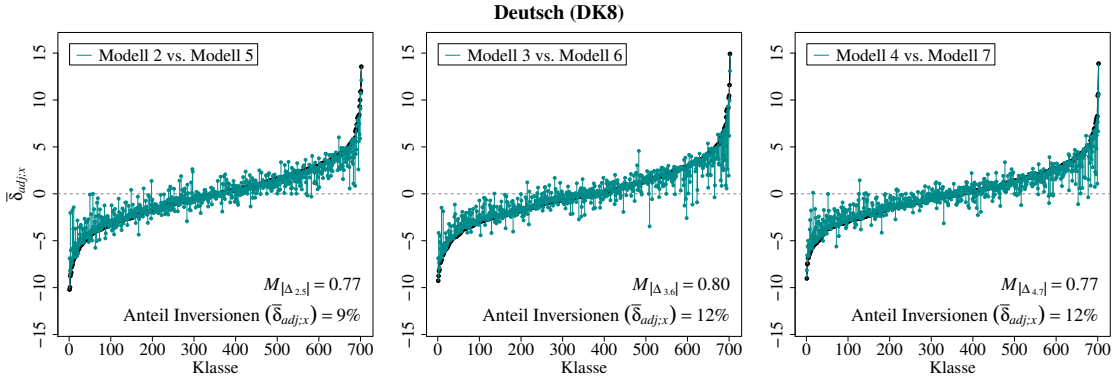

Abbildung 7.26: Change-Plots im Fach Deutsch (DK8): VAM *versus* CVA (bedingt lineare Parametrisierung mit Interaktionen)

ker ist die Veränderung infolge der Hinzunahme des fachspezifischen Vorwissens in Klassenstufe 6 und der entsprechenden Kompositionsmerkmale (Modell 3 vs. Modell 6): Hier beträgt die durchschnittliche Differenz der Effektschätzungen $M_{|\Delta_{3.6}|} = 0.80$. Zudem finden sich bei 12% der Klassen ($n = 83$) Inversionen der Effekte beim Wechsel von Modell 3 zu Modell 6. Werden sowohl DK3 als auch DK6 sowie die entsprechende leistungsmäßige Klassenkomposition in das Modell aufgenommen (Modell 4 vs. Modell 7), liegt die durchschnittliche Differenz der Effektschätzungen bei $M_{|\Delta_{4.7}|} = 0.77$. Jedoch ist auch hier für 12% der Klassen ($n = 85$) eine Inversion der Effekte beobachtbar.

(3) *Bedingte Unabhängigkeit*:
Ist es im Fachbereich Deutsch hinreichend, statt beider Vorwissensvariablen (DK3

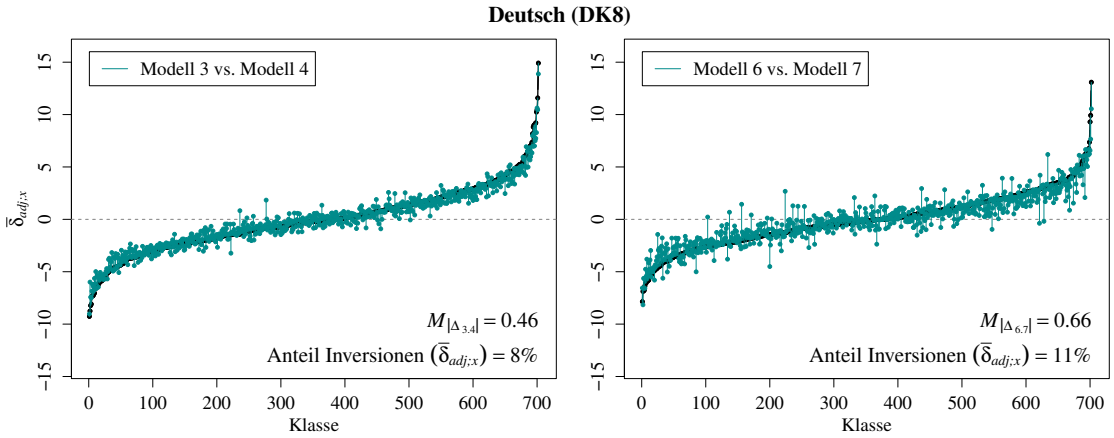

Abbildung 7.27: Change-Plots im Fach Deutsch (DK8): Modelle mit *versus* ohne DK3 (bedingt lineare Parametrisierung mit Interaktionen)

und DK6) lediglich das fachspezifische Vorwissen aus Klassenstufe 6 (DK6) zusätzlich in das Modell aufzunehmen? Bei dieser Frage geht es erneut um die bedingte Unabhängigkeit der Testwertvariablen DK8 von DK3 gegeben DK6 und der restlichen Kovariaten im Adjustierungsmodell. Wäre dies der Fall, so sollte der Wechsel von Modell 3 zu Modell 4 zu keinen Veränderungen in den adjustierten Effektschätzungen führen. Der entsprechende Vergleich der Effektschätzungen aus beiden Modellen ist auf der linken Seite von Abbildung 7.27 wiedergegeben. Das Ergebnis spricht m. E. gegen die entsprechende bedingte Unabhängigkeitsannahme, da der Modellwechsel mit deutlich stärkeren Veränderungen einhergeht als beim entsprechenden Vergleich im Fach Mathematik: Die durchschnittliche Abweichung liegt hier bei $M_{|\Delta_{3.4}|} = 0.46$ mit einem 8%-igem Anteil an Inversionen ($n = 57$ Klassen). Ein diesbezüglich noch deutlicheres Ergebnismuster zeigt sich bei Betrachtung der Sensitivität der Effektschätzungen bei Hinzunahme der entsprechenden Kompositionsmerkmale: Die rechten Grafik in Abbildung 7.27 stellt den Vergleich der Effektschätzungen aus den Modellen 6 und 7 dar. Die durchschnittliche Veränderung der Effektschätzungen liegt bei $M_{|\Delta_{6.7}|} = 0.66$, wobei sich bei 11% ($n = 77$) der Klassen die Richtung des Effekts (positiv vs. negativ) umkehrt.

Vergleicht man diese Ergebnisse zwischen den Fachbereichen, so zeigt sich, dass zwar die mittleren Beträge der Veränderungen in den Effektschätzungen jeweils deutlich größer im Fach Deutsch als im Fach Mathematik sind. Dies wird auch in den einzelnen

Change-Plots ersichtlich, die im Fach Deutsch wesentlich „unruhiger“ sind als in Mathematik. Hinsichtlich des Anteils an Inversionen der adjustierten Effektschätzungen infolge des Wechsels des Analysemodells zeigt sich jedoch zwischen den Fächern ein sehr ähnliches Bild – sowohl im Muster der Ergebnisse über die verschiedenen paarweisen Modellvergleiche hinweg als auch in der absoluten Höhe des entsprechenden Prozentsatzes.

Lineare Parametrisierung (ohne Interaktionen). Kommen wir nun zu den Modellen mit linearer Parametrisierung ohne Interaktionen (Modelle 8 bis 14), deren Ergebnisse nachfolgend einer vergleichenden Betrachtung unterzogen werden.

(1) *CAM vs. VAM*:
Abbildung 7.28 zeigt die drei Change-Plots, die aus dem Wechsel des Adjustierungmodells von Modell 8 (CAM) zu je einem Modell des Typs VAM (Modell 9, 10 bzw. 11) resultieren.
Der linke Change-Plot zeigt die Veränderungen der adjustierten klassenspezifischen Effektschätzungen infolge der zusätzlichen Berücksichtigung von DK3 (Modell 8 vs. Modell 9). Der Mittelwert des Betrages dieser Veränderungen ist $M_{|\Delta_{8.9}|} = 1.08$, wobei der Anteil der Inversionen bei 10% ($n = 69$) liegt. Wird hingegen DK6 anstatt DK3 zusätzlich in dem Adjustierungsmodell berücksichtigt (Modell 8 vs. Modell 10), so beträgt die durchschnittliche Abweichung $M_{|\Delta_{8.10}|} = 1.69$. Zudem steigt – wie auch bei den bedingt linear parametrisierten Modellen (vgl. Abbildung 7.25) – der Anteil der Inversionen in $\bar{\delta}_{adj;x}$ auf 16% und betrifft somit $n = 115$ der insgesamt $N = 702$ betrachteten Klassen. Ein ähnliches Bild zeigt sich, wenn sowohl DK3 als auch DK6 zusätzlich in das Modell aufgenommen werden (Modell 8 vs. Modell 11). Dabei resultiert eine durchschnittliche Veränderung von $M_{|\Delta_{8.11}|} = 1.79$ und der Anteil der Inversionen beträgt sogar 18% ($n = 125$ Klassen).

(2) *VAM vs. CVA*:
Welchen Einfluss hat die zusätzliche Berücksichtigung der leistungsmäßigen Klassenkomposition auf die Veränderungen der adjustierten Effektschätzungen in den linearen Adjustierungsmodellen im Fach Deutsch? Die Ergebnisse der diese Frage adressierenden Modellvergleiche (Modell 9 vs. 12, Modell 10 vs. 13 sowie Modell 11 vs. 14) sind in Abbildung 7.29 dargestellt.

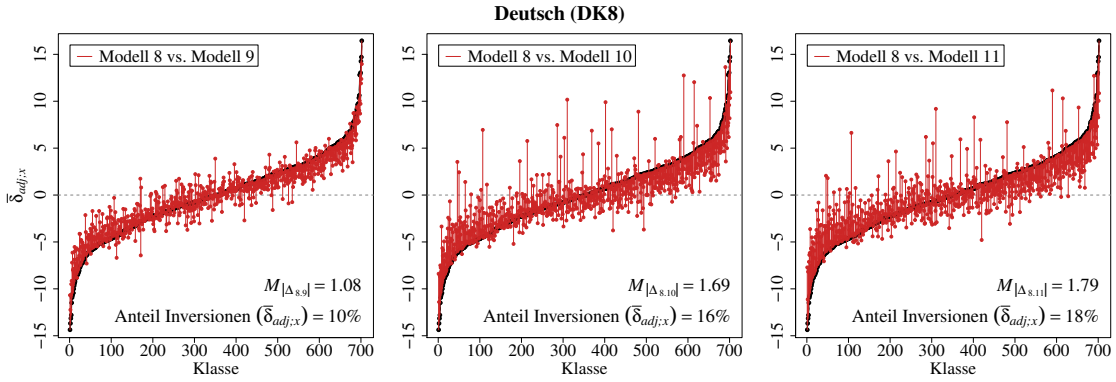

Abbildung 7.28: Change-Plots im Fach Deutsch (DK8): CAM *versus* VAM (lineare Parametrisierung ohne Interaktionen)

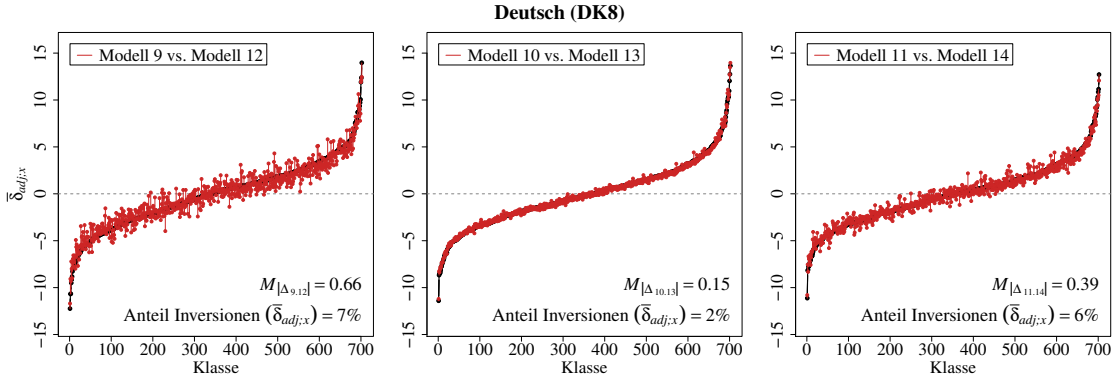

Abbildung 7.29: Change-Plots im Fach Deutsch (DK8): VAM *versus* CVA (lineare Parametrisierung ohne Interaktionen)

Werden zusätzlich zu den restlichen Kovariaten *und* zum fachspezifischen Vorwissen aus Klassenstufe 3 (DK3) auch die entsprechenden Klassenkompositionsmerkmale hinsichtlich des Vorwissens aus Klassenstufe 3 in das Modell aufgenommen (Modell 9 vs. Modell 12), so liegt der Mittelwert des Betrages der Differenzen zwischen den Effektschätzungen bei $M_{|\Delta_{9.12}|} = 0.66$. Gleichzeitig wird bei 7% aller Klassen ($n = 47$) ein positiver zu einem negativen Effekt und umgekehrt. Deutlich geringer ist diese Veränderung infolge der Hinzunahme des fachspezifischen Vorwissens in Klassenstufe 6 und der entsprechenden Kompositionsmerkmale (Modell 10 vs. Modell 13): Hier beträgt die durchschnittliche Differenz der Effektschätzungen $M_{|\Delta_{10.13}|} = 0.15$. Zudem finden sich bei lediglich 2% der Klassen ($n = 16$) Inversionen der Effekte beim Wechsel von Modell 3 zu Modell 6. Und auch für den Fall, dass Informationen sowohl zum fachspe-

zifischen Vorwissen aus Klassenstufe 3 (DK3) als auch Klassenstufe 6 (DK6) sowie der entsprechenden Klassenkomposition zur Verfügung stehen (Modell 11 vs. 14), ist die durchschnittliche Differenz der Effektschätzungen recht gering ($M_{|\Delta_{11.14}|} = 0.39$), wobei hier hingegen für etwa $n = 39$ Klassen (6%) eine Inversion der Effekte beobachtbar ist.

Zusammenfassend zeigt sich auch bei der linearen Parametrisierung – wie bereits beim entsprechenden Vergleich der bedingt linearen Adjustierungsmodelle – eine geringere Sensitivität der Effektschätzungen als in Abbildung 7.28: Die Veränderungen infolge der zusätzlichen Berücksichtigung von Klassenkompositionsmerkmalen sind jeweils geringer als bei der Hinzunahme des fachspezifischen Vorwissens. Anders als im Fachbereich Mathematik gibt es hier jedoch deutliche Unterschiede hinsichtlich der Sensitivität in Abhängigkeit von der Kovariatenselektion: Je nachdem, ob die leistungsmäßige Klassenkomposition basierend auf DK3, DK6 *oder* beiden (DK3 und DK6) zusätzlich im linearen Adjustierungsmodell berücksichtigt wird, zeigt sich eine unterschiedlich starke Sensitivität der adjustierten klassenspezifischen Effektschätzungen. Diese ist am geringsten beim Wechsel von Modell 10 zu 11, d. h. bei Hinzunahme von DK6 und entsprechender Klassenkompositionsmerkmale.

(3) *Bedingte Unabhängigkeit*:
Hinsichtlich der bedingten Unabhängigkeit der Testwertvariablen DK8 von DK3 gegeben DK6 und der restlichen Kovariaten im linearen Adjustierungsmodell (ohne Interaktionen) zeigt sich ein ähnliches Ergebnis wie bereits bei der bedingt linearen Parametrisierung (vgl. Abbildung 7.27): Der entsprechende Vergleich der Effektschätzungen aus den Modellen 10 und 11 ist auf der linken Seite von Abbildung 7.30 wiedergegeben. Das Ergebnis spricht gleichfalls gegen die bedingte Unabhängigkeitsannahme, da der Modellwechsel mit deutlichen Veränderungen einhergeht: Die durchschnittliche Abweichung liegt hier bei $M_{|\Delta_{10.11}|} = 0.44$ mit einem 6%-igem Anteil an Inversionen ($n = 44$ Klassen). Ein diesbezüglich noch deutlicheres Ergebnismuster zeigt sich bei Betrachtung der Sensitivität der Effektschätzungen bei Hinzunahme der entsprechenden Kompositionsmerkmale: Die rechte Grafik in Abbildung 7.30 stellt den Vergleich der Effektschätzungen aus den Modellen 13 und 14 dar. Die durchschnittliche Veränderung der Effektschätzungen ist $M_{|\Delta_{13.14}|} = 0.79$, wobei sich bei 10% der Klassen

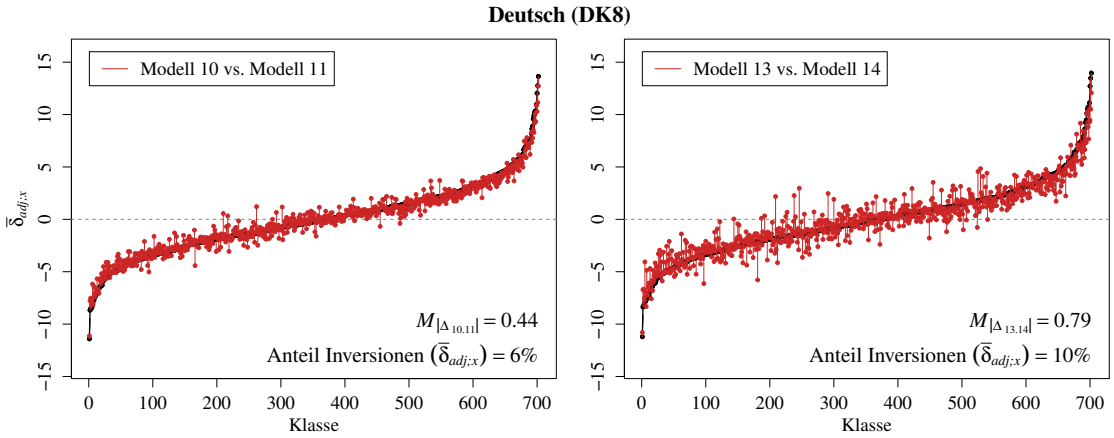

Abbildung 7.30: Change-Plots im Fach Deutsch (DK8): Modelle mit *versus* ohne DK3 (lineare Parametrisierung ohne Interaktionen)

($n = 73$) die Richtung des Effekts (positiv oder negativ) umkehrt.

In Relation zum zuvor dargestellten Vergleich der Modelle mit saturierter und bedingt linearer Parametrisierung (mit Interaktionen) fällt auf, dass sich zwar die durchschnittlichen Beträge der Abweichungen der Effektschätzungen beim Wechsel der Modelle unterscheiden. Beim Wechsel vom CAM zum VAM sind diese stets größer beim Vergleich linear parametrisierter Modelle. Jedoch ist der Anteil der Klassen, bei denen sich die adjustierten Effektschätzungen infolge des Modellwechsels von einem positiven in einen negativen Effekt (und vice versa) umkehrt, jeweils fast identisch. So ist bspw. die durchschnittliche Abweichung der Effekte $M_{|\Delta_{1.3}|} = 1.43$, wenn DK6 zusätzlich zu den anderen Kovariaten in das Modell mit bedingt linearer Parametrisierung aufgenommen wird (vgl. Abbildung 7.25). Die Abweichung zwischen den beiden hinsichtlich der Kovariatenselektion einander entsprechenden Modellen mit linearer Parametrisierung beträgt hingegen $M_{|\Delta_{8.10}|} = 1.69$ (vgl. Abbildung 7.28). Jedoch ist der Anteil der Inversionen mit 17% bzw. 16% ($n = 117$ bzw. $n = 115$ der insgesamt $N = 702$ betrachteten Klassen) nahezu gleich.

Anders als im Fach Mathematik zeigt sich das soeben dargestellte Ergebnismuster hingegen nicht beim Wechsel vom VAM zum CVA. Hier sind sowohl die mittleren Beträge der Abweichungen als auch der Anteil an Inversionen in den Modellen mit linearer Parametrisierung deutlich geringer: Während bspw. die durchschnittliche Differenz der Effekte beim Vergleich der Modelle 3 und 6 bei $M_{|\Delta_{3.6}|} = 0.80$ liegt (vgl. Abbildung 7.26), beträgt diese bei hinsichtlich des Kovariatensets äquivalenten Modellen mit li-

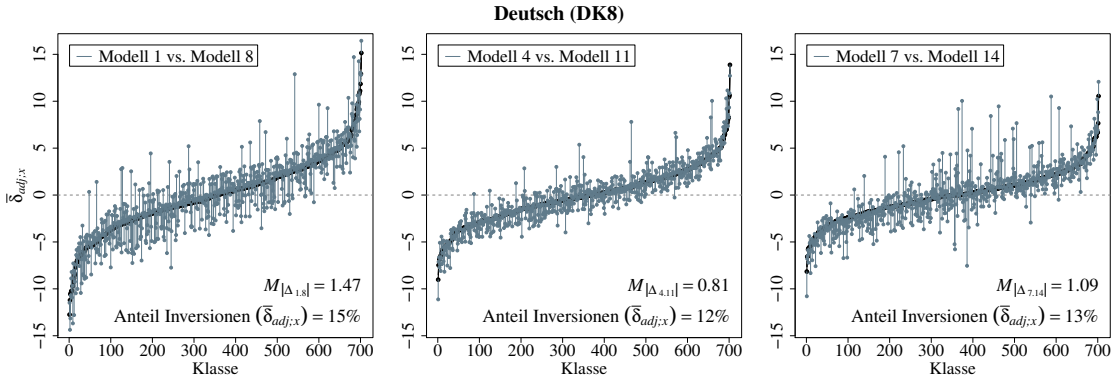

Abbildung 7.31: Change-Plots im Fach Deutsch (DK8): Bedingt lineare *versus* lineare Parametrisierung

nearer Parametrisierung lediglich $M_{|\Delta_{10.13}|} = 0.15$ (vgl. Abbildung 7.29). Gleichzeitig sinkt der Anteil der Inversionen von 12% auf 2% der Klassen.

Bedingt lineare vs. lineare Parametrisierung. Schließlich werden – analog zu den Analysen im Fach Mathematik – auch die verschiedenen Parametrisierungen einer vergleichenden Betrachtung unterzogen, um die Sensitivität der adjustierten Effektschätzungen infolge der Modifikation der Modellselektion beurteilen zu können.

Abbildung 7.31 zeigt die Veränderungen der adjustierten klassenspezifischen Effektschätzungen infolge des Wechsels der Modellspezifikation von einer saturierten bzw. bedingt linearen zu einer linearen Parametrisierung. Welchen Einfluss hat die Modellselektion auf die adjustierten klassenspezifischen Effektschätzungen bei dem CAM (Modell 1 vs. Modell 8)? Die durchschnittliche Veränderung der Effektschätzungen beträgt hier $M_{|\Delta_{1.8}|} = 1.47$ und für 15% der Klassen ($n = 105$) kehrt sich die Richtung des Effekts um. Beim Vergleich der zwei VAM, die beide das fachspezifische Vorwissen der Jahrgangsstufen 3 und 6 zusätzlich zu den Kovariaten im CAM enthalten (Modell 4 vs. Modell 11), ist diese Veränderung hingegen geringer: Der Mittelwert des Betrages dieser Veränderungen liegt bei $M_{|\Delta_{4.11}|} = 0.81$. Auch der Anteil der Inversionen ist hier mit 12% bzw. $n = 84$ Klassen geringer als beim vorangegangen Modellvergleich. Schließlich ergibt ein Vergleich der zwei CVA, die sich hinsichtlich der Parametrisierung unterscheiden (Modell 7 vs. Modell 14) eine durchschnittliche Veränderung der Effektschätzungen von $M_{|\Delta_{7.14}|} = 1.09$. Der Anteil der Inversionen beträgt 13%, d. h. $n = 94$ der insgesamt $N = 702$ Klassen.

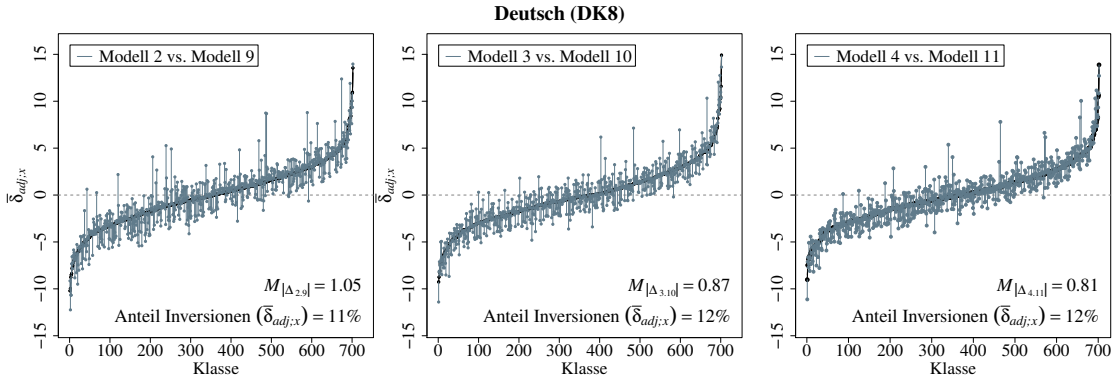

Abbildung 7.32: Change-Plots im Fach Deutsch (DK8): VAM mit bedingt linearer *versus* linearer Parametrisierung

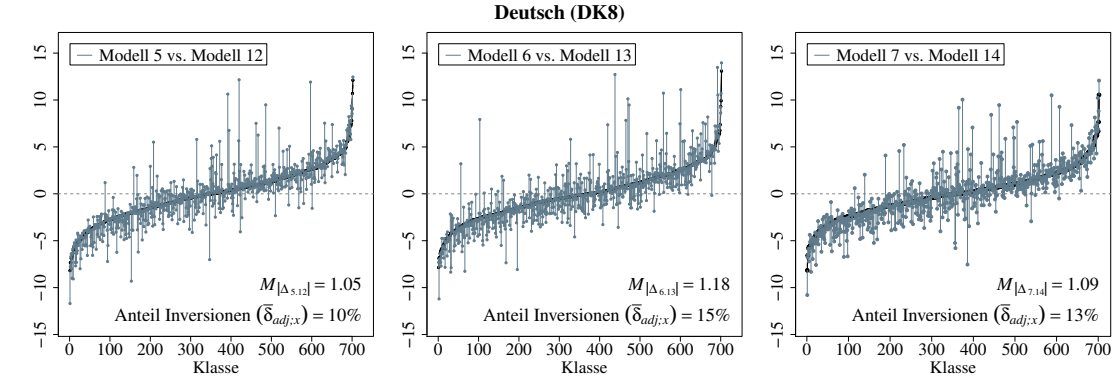

Abbildung 7.33: Change-Plots im Fach Deutsch (DK8): CVA mit bedingt linearer *versus* linearer Parametrisierung

Wie bereits in Mathematik ist auch für Deutsch der Einfluss der Parametrisierung insgesamt am stärksten für die CAM (d. h. Modelle ohne das fachspezifische Vorwissen). Dieser Einfluss nimmt ab, wenn das fachspezifische Vorwissen im Modell enthalten ist (VAM) bzw. auch die leistungsmäßige Klassenkomposition berücksichtigt wird (CVA). Jedoch sind die Unterschiede zwischen den drei Modellvergleichen in Abbildung 7.31 im Fach Deutsch geringer als im Fach Mathematik. Vergleicht man VAM und CVA, so ist erneut eine höhere Stabilität der VAM beobachtbar: So führt der Parametrisierungswechsel von Modell 4 zu Modell 11 (VAM) zu durchschnittlich geringeren Unterschieden der Effektschätzungen *und* zu einem kleineren Anteil an Inversionen als der entsprechenden Wechsel von Modell 7 zu Modell 14 (CVA). Diese größere Stabilität der VAM gegenüber des Wechsels der Parametrisierung hinsichtlich der Veränderung der Effektschätzungen einzelner Klassen ist somit auch im Fach Deutsch konkordant

mit dem Ergebnismuster hinsichtlich der Korrelationen (vgl. Abschnitt 7.3.3).

Zudem zeigt sich jedoch gleichfalls eine deutliche Variabilität – in Bezug auf die Sensitivität der Effektschätzungen – *zwischen* den einzelnen VAM respektive CVA. Die zugehörige grafische Darstellung dieser Modellvergleiche findet sich in den Abbildungen 7.32 und 7.33. Innerhalb der Modellklasse VAM schwanken die durchschnittlichen Abweichungen lediglich zwischen minimal $M_{|\Delta_{4.11}|} = 0.81$ und maximal $M_{|\Delta_{2.9}|} = 1.05$. Der Anteil der Inversionen beträgt zwischen 12% ($n = 84$) und 11% ($n = 81$) der Klassen. Innerhalb der Modellklasse CVA hingegen variieren die durchschnittlichen Abweichungen zwischen maximal $M_{|\Delta_{6.13}|} = 1.18$ und minimal $M_{|\Delta_{5.12}|} = 1.05$, wobei gleichzeitig zwischen 15% ($n = 106$) und 10% ($n = 74$) der Klassen von einer Inversion der Effektschätzung betroffen sind.

Zusammenfassung: Change-Plots

Da die Ergebnisse aus Vergleichsarbeiten auf Klassenebene ausgewertet und zurückgemeldet werden, ist insbesondere die Veränderung der Effektschätzungen einzelner Klassen relevant. Um diesbezüglich die Sensitivität der Effektschätzungen für einzelne Klassen quantifizieren und beurteilen zu können, eignen sich im Besonderen Change-Plots. Für diese Darstellungsform zeigt sich durchgängig eine recht hohe Sensitivität der adjustierten klassenspezifischen Effektschätzungen gegenüber Modifikationen der Kovariaten- und Modellselektion. Wird das fachspezifische Vorwissen bzw. die leistungsmäßige Klassenkomposition zusätzlich in das Adjustierungsmodell aufgenommen oder aber die Parametrisierung geändert, finden sich insgesamt recht „unruhige“ Change-Plots. So liegt u. a. der Anteil der Klassen, deren Effekte sich von einem positiven in einen negativen Effekt ändert (und umgekehrt), stets über 5%. Das Ausmaß und die Richtung des adjustierten Effekts einer Klasse ist somit in deutlichem Maße abhängig von der Wahl der Kovariaten und der gewählten Parametrisierung.

Im Hinblick auf die Kovariatenselektion zeigt sich – für jeweils beide der Parametrisierungsformen – beim Wechsel von CAM zum VAM eine deutliche Sensitivität der Effektschätzungen. Dies spricht für Hypothese 1.1 über den Einfluss des fachspezifischen Vorwissens. Zwar sind diese Veränderungen – sowohl hinsichtlich des Ausmaßes (Mittelwert der Beträge der Differenzen der adjustierten Effekte) als auch hinsichtlich der Richtung (Anteil der Inversionen) – Wechsel vom VAM zum CVA geringer, jedoch beträgt der Anteil der Inversionen stets ca. 10%. Dieser Befund stützt gleichsam Hy-

pothese 1.2. Zwischen den drei verschiedenen VAM-CVA-Vergleichen sind diese Veränderungen zudem homogener (d. h. ähnlich groß und häufig) als zwischen den CAM-VAM-Vergleichen. Insbesondere zeigt sich beim Wechsel vom CAM zum VAM eine deutliche höhere Sensitivität der Effektschätzungen, wenn das fachspezifische Vorwissen aus Klassenstufe 6 (anstatt aus Klassenstufe 3) in das Modell aufgenommen wird.

Des Weiteren ergibt der Vergleich zwischen den beiden Parametrisierungsformen: (a) Ersetzt man – gegeben einem konkreten Kovariatenset – die bedingt lineare durch eine lineare Parametrisierung, so finden sich deutliche Veränderungen der klassenspezifischen Effektschätzungen. Dies spricht für Hypothese 2. (b) Gemäß Hypothese 3 sollten die Veränderungen der klassenspezifischen Effektschätzungen infolge des Wechsels der Parametrisierung des Adjustierungsmodells über die drei Modelltypen CAM, VAM und CVA abnehmen. Zwar zeigen sich die stärksten Veränderungen zwischen den CAM, die beim Wechsel zum VAM (also infolge der Hinzunahme des fachspezifischen Vorwissens) abnehmen. Jedoch ergibt sich sowohl für den durchschnittlichen Betrag dieser Veränderungen als auch im Anteil an Inversionen tendenziell eine stärkere Sensitivität beim Wechsel zum CVA, was wiederum gegen Hypothese 3 spricht.

Schließlich findet sich auch bei den Change-Plots das in Hypothese 4 postulierte konkordante Ergebnismuster in beiden Fachbereichen.

7.3.5 Transitionsmatrizen

Um die Sensitivität der adjustierten klassenspezifischen Effektschätzungen gegenüber der Kovariaten- und Modellselektion zu verdeutlichen, verwende ich nachfolgend ein weiteres Kriterium: die Veränderungen des Quintil-Rankings der klassenspezifischen Effektschätzungen infolge des Wechsels von einem zu einem anderen Modell (z. B. Benton et al., 2003; Goldhaber, Goldschmidt & Tseng, 2013). Dazu werden die adjustierten klassenspezifischen Effektschätzungen pro Modell zunächst in Quintile aufgeteilt: Quintil 1 enthält 20% der Klassen mit dem größten positiven adjustierten Effekt, wohingegen Quintil 5 jeweils 20% aller Klassen mit dem größten negativen adjustierten Effekt enthält. Die nachfolgenden Tabellen stellen schließlich die Veränderungen dieses Quintil-Rankings der adjustierten Effektschätzungen beim Wechsel von einem zu einem anderen Adjustierungsmodell dar. Diese Veränderungen werden in einer sog. *Transitionsmatrix* (engl.: transition matrix; vgl. Goldhaber et al., 2013, S. 10) wiederum für einen paarweisen Modellvergleich abgebildet: Die Zellen der Transitionsmatrix

enthalten den prozentualen Anteil (zeilenweise) an Klassen, die in ein konkretes Quintil basierend auf einem ersten Modell fallen und in das gleiche oder ein anderes Quintil, wenn ein zweites Modell als Grundlage der Berechnung dient.

Ein Beispiel soll dies verdeutlichen: In Tabelle 7.11 (obere Transitionsmatrix) wird der Modellvergleich von Modell 1 mit Modell 4 dargestellt. Die Zeilen repräsentieren die Quintile der Effektschätzungen aus Modell 1 und die Spalten repräsentieren die Quintile der Effektschätzungen aus Modell 4. Bei 73.10% der Thüringer Klassen im Fach Mathematik, die gemäß Modell 1 in das erste Quintil eingeordnet werden, sind die adjustierten Effektschätzungen auch aus Modell 4 (VAM) dem ersten Quintil zuzuordnen. Für die restlichen 26.90% der Klassen, die basierend auf Modell 1 dem ersten Quintil zugeordnet werden, findet eine Rangverschlechterung statt, falls stattdessen Modell 4 angewendet wird: 22.07% der Klassen fallen dann in das zweite Quintil, 4.14% in das dritte und lediglich 0.69% in das vierte Quintil. Keine Klasse, die gemäß Modell 1 dem ersten Quintil zuzuordnen ist, fällt beim Wechsel zu Modell 4 in das fünfte Quintil.

Je größer die Prozentsätze in den Zellen der Diagonale (d. h. je näher an 100%) und entsprechend je kleiner diese in den Zellen außerhalb der Diagonale sind, desto stabiler sind die Effektschätzungen beim Wechsel des Adjustierungsmodells. Dies wäre ein Indikator einer geringen Sensitivität der adjustierten klassenspezifischen Effektschätzungen. Falls die Wahl des Adjustierungsmodells irrelevant für die Zuordnung der Klassen – basierend auf den resultierenden Effektschätzungen – zu einem der Quintile ist, werden beide Modelle zu identischen Zuordnungen für alle Klassen führen. Dann wäre einzig die Diagonale besetzt (mit je 100% pro Zeile) und die Elemente der oberen und unteren Dreiecksmatrix (oberhalb bzw. unterhalb der Diagonale) würden ausschließlich den Wert null annehmen.

Der Vorteil dieses Kriteriums besteht darin, dass Veränderungen der adjustierten Effektschätzungen auf diese Weise sehr übersichtlich dargestellt werden können. Zudem erlauben diese eine zusätzliche Einschätzung der Bedeutsamkeit bzw. Stärke der Veränderungen. Der bei den Change-Plots (Abschnitt 7.3.4) angegebene Anteil der Inversionen bezüglich der klassenspezifischen Effektschätzungen wird auf diese Weise um die Information ergänzt, wie stark diese Inversion ist: Wird bspw. eine Klasse auf Basis eines Modells dem zweiten Quintil, basierend auf einem anderen Modell jedoch dem vierten Quintil zugeordnet, so liegt eine Inversion des Effekts vor. Diese ist jedoch umso stärker, je mehr Quintile „übersprungen“ werden. Demnach ist die maximale Transiti-

Tabelle 7.11: Veränderungen des Quintil-Rankings der adjustierten Effektschätzungen im Fach Mathematik (MK8) infolge der zusätzlichen Berücksichtigung des fachspezifischen Vorwissens: CAM *versus* VAM

Bedingt lineare Parametrisierung (inkl. Interaktionen):						
		Modell 4				
		1	2	3	4	5
Modell 1	1	73.10	22.07	4.14	0.69	0
	2	21.38	39.31	31.03	7.59	0.69
	3	4.17	27.78	33.33	29.86	4.86
	4	1.38	10.34	26.90	40.69	20.69
	5	0	0.69	4.14	21.38	73.79
Lineare Parametrisierung (ohne Interaktionen):						
		Modell 11				
Modell 8	1	71.72	25.52	2.76	0	0
	2	25.52	40.69	23.45	8.28	2.07
	3	2.08	25.69	40.28	26.39	5.56
	4	0.69	6.90	25.52	44.14	22.76
	5	0	1.38	7.59	21.38	69.66

Anmerkungen. Die Werte in den Zellen der Tabelle sind Prozent der Klassen pro Zeile, deren Effektschätzung im entsprechenden Quintilsabschnitt liegt. Die zeilenweise Summe beträgt somit jeweils 100%. Anzahl der Klassen: $N = 724$. Anzahl der Klassen pro Zeile bzw. pro Spalte: $n_1 = 145$, $n_2 = 145$, $n_3 = 144$, $n_4 = 145$, $n_5 = 145$.

on eine Veränderung um vier Quintile, d. h., wenn eine Klasse auf Basis eines Modells dem ersten Quintil, basierend auf einem anderen Modell jedoch dem fünften Quintil zugeordnet wird (bzw. vice versa).

Im Folgenden werde ich zunächst die Ergebnisse für das Fach Mathematik und anschließend für das Fach Deutsch darstellen.

Transitionsmatrizen im Fach Mathematik

Kovariatenselektion. Entsprechend der bisherigen Ergebnisdarstellung werde ich zunächst Modelle mit jeweils identischer Parametrisierung vergleichen, die sich ausschließlich in der Wahl der Kovariaten unterscheiden.

(1) *CAM vs. VAM*:

Welchen Einfluss auf das Quintil-Ranking der Klassen hat die zusätzliche Berücksichtigung des fachspezifischen Vorwissens (MK3 und MK6)? Tabelle 7.11 zeigt die Transitionsmatrizen, die beim Wechsel vom CAM zum VAM resultieren[15]. Die obere Matrix in Tabelle zeigt die Transitionen für Modelle mit bedingt linearer Parametrisierung inklusive Interaktionen (Modell 1 vs. Modell 4). Bei Betrachtung der Diagonale fällt auf, dass der Modellwechsel mit erheblichen Veränderungen des Quintil-Rankings einhergehen – insbesondere bei den mittleren Quintilen 2, 3 und 4, bei denen jeweils nur 39.31%, 33.33% und 40.69% der Klassen infolge des Modellwechsels in das gleiche Quintil fallen. Am stabilsten sind die beiden extremen Quintile 1 und 5 mit 73.10% bzw. 73.79%. Die untere Matrix hingegen zeigt die Transitionen für die hinsichtlich der Kovariaten entsprechenden Modelle mit linearer Parametrisierung, in denen potenzielle Interaktionen der Kovariaten nicht modelliert sind (Modell 8 vs. Modell 11). Insgesamt zeigt sich hier ein zu dem in der oberen Matrix sehr ähnliches Ergebnismuster. Die Stabilität in den drei mittleren Quintilen 2, 3 und 4 beträgt jeweils 40.69%, 40.28% bzw. 44.14%. Hingegen sind die Extrempole der Diagonale – Quintil 1 bzw. Quintil 5 mit 71.72% respektive 69.66% – wiederum am stärksten besetzt.

(2) *VAM vs. CVA*:

Tabelle 7.12 zeigt die resultierenden Transitionsmatrizen, die beim Wechsel vom VAM zum CVA resultieren. Dieser Vergleich adressiert die Sensitivität der Effektschätzungen hinsichtlich der Quintil-Verteilung infolge der zusätzlichen Berücksichtigung der leistungsmäßigen Klassenkomposition. Wiederum in der oberen Matrix von Tabelle 7.12 sind die Transitionen bei bedingt linearer Parametrisierung (mit Interaktionen) dargestellt, wobei hier Modell 4 und Modell 7 verglichen werden. Zwar zeigen sich auch hier deutliche Veränderungen des Quintil-Rankings der Klassen basierend auf den jeweils resultierenden Effektschätzungen. Jedoch sind diese Veränderungen im Vergleich zum Wechsel vom CAM zum VAM geringer. Mit anderen Worten: Die Effektschätzungen sind beim Wech-

[15] Das im Rahmen dieser Arbeit verwendete Design des Modellvergleichs erlaubt – bei konstanter Parametrisierung der Modelle – insgesamt jeweils drei Vergleiche beim Wechsel vom CAM zum VAM und ebenso beim Wechsel vom VAM zum CVA. Aus Gründen der Übersichtlichkeit der Ergebnisdarstellung wird nachfolgend immer der Vergleich mit dem hinsichtlich der Kovariaten jeweils komplexesten Modell dargestellt. Die hier nicht abgebildeten Vergleiche zeigen ein konkordantes Ergebnismuster (vgl. Anhang F).

Tabelle 7.12: Veränderungen des Quintil-Rankings der adjustierten Effektschätzungen im Fach Mathematik (MK8) infolge der zusätzlichen Berücksichtigung von Kompositionsmerkmalen: VAM *versus* CVA

Bedingt lineare Parametrisierung (inkl. Interaktionen):						
		Modell 7				
		1	2	3	4	5
Modell 4	1	75.86	19.31	4.14	0.69	0
	2	22.07	56.55	17.93	3.45	0
	3	2.08	21.53	55.56	20.14	0.69
	4	0	2.07	19.31	63.45	15.17
	5	0	0.69	2.76	12.41	84.14
Lineare Parametrisierung (ohne Interaktionen):						
		Modell 14				
Modell 11	1	82.76	16.55	0.69	0	0
	2	16.55	54.48	24.83	4.14	0
	3	0.69	27.78	54.17	16.67	0.69
	4	0	1.38	20.00	62.07	16.55
	5	0	0	0	17.24	82.76

Anmerkungen. Die Werte in den Zellen der Tabelle sind Prozent der Klassen pro Zeile, deren Effektschätzung im entsprechenden Quintilsabschnitt liegt. Die zeilenweise Summe beträgt somit jeweils 100%. Anzahl der Klassen: $N = 724$. Anzahl der Klassen pro Zeile bzw. pro Spalte: $n_1 = 145$, $n_2 = 145$, $n_3 = 144$, $n_4 = 145$, $n_5 = 145$.

sel vom VAM zum CVA stabiler. So betragen die Werte der Diagonale, welche den Prozentsatz der nach beiden Modellen identisch kategorisierten Klassen indizieren, zwischen minimal 55.56% (Quintil 3) und maximal 84.14% (Quintil 5). Wie auch beim vorangegangenen Vergleich (CAM vs. VAM) sind hier die beiden Quintil-Pole am stabilsten (75.86% in Quintil 1 und 84.14% in Quintil 5). Die untere Matrix in Tabelle 7.12 zeigt die Transitionen für die hinsichtlich der Kovariaten entsprechenden Modelle mit linearer Parametrisierung, in denen potenzielle Interaktionen der Kovariaten nicht modelliert sind (Modell 11 vs. Modell 14). Hier zeigt sich ein insgesamt vergleichbares Ergebnismuster wie in der oberen Matrix. So sind bspw. auch hier die beiden Quintil-Pole am stabilsten mit

Tabelle 7.13: Veränderungen des Quintil-Rankings der adjustierten Effektschätzungen im Fach Mathematik (MK8) bei bedingt linearer Parametrisierung: Bedingte Unabhängigkeit

Bedingt lineare Parametrisierung (inkl. Interaktionen):						
		Modell 4				
		1	2	3	4	5
Modell 3	1	93.79	6.21	0	0	0
	2	6.21	83.45	10.34	0	0
	3	0	10.42	79.17	10.42	0
	4	0	0	10.34	84.14	5.52
	5	0	0	0	5.52	94.48
Lineare Parametrisierung (ohne Interaktionen):						
				Modell 7		
Modell 6	1	82.76	15.86	1.38	0	0
	2	17.24	66.21	15.17	1.38	0
	3	0	17.36	63.19	19.44	0
	4	0	0.69	20.00	66.21	13.10
	5	0	0	0	13.10	86.90

Anmerkungen. Die Werte in den Zellen der Tabelle sind Prozent der Klassen pro Zeile, deren Effektschätzung im entsprechenden Quintilsabschnitt liegt. Die zeilenweise Summe beträgt somit jeweils 100%. Anzahl der Klassen: $N = 724$. Anzahl der Klassen pro Zeile bzw. pro Spalte: $n_1 = 145$, $n_2 = 145$, $n_3 = 144$, $n_4 = 145$, $n_5 = 145$.

82.76% in Quintil 1 und 82.76% in Quintil 5.

(3) *Bedingte Unabhängigkeit*:
Führt die zusätzliche Berücksichtigung beider Vorwissensvariablen (MK3 und MK6) – im Vergleich zur Hinzunahme von MK6 allein – zu substanziellen Veränderungen im Quintil-Ranking der Klassen? Zu diesem Zweck vergleichen wir die Ergebnisse aus den Modellen 3 vs. 4 (bedingt lineare Parametrisierung mit Interaktionen), die in der oberen Matrix von Tabelle 7.13 wiedergegeben sind. Hier zeigt sich – in Relation zu den bisherigen Vergleichen – die stärkste Stabilität der Effektschätzungen: Die Werte der Diagonale, die den Prozentsatz der nach beiden Modellen identisch kategorisierten Klassen indizieren, betragen 93.79%

in Quintil 1, 83.45% in Quintil 2, 79.17% in Quintil 3, 84.14% in Quintil 4 und 94.48% in Quintil 5. Des Weiteren sind 12 der 25 Zellen in der Transitionsmatrix unbesetzt (je 0% der Klassen) und Veränderungen des Quintil-Rankings zu lediglich benachbarten Quintilen (z. B. von Quintil 3 zu 2 bzw. von 3 zu 4) beobachtbar. Eine demgegenüber geringere Stabilität zeigt sich, wenn Klassenkompositionsmerkmale aus den jeweiligen Klassenstufen – zusätzlich zu den restlichen Kovariaten und den Vorwissensvariablen – berücksichtigt werden. Die entsprechenden Ergebnisse finden sich in der unteren Transitionsmatrix von Tabelle 7.13. Hier betragen die Werte der Diagonale zwischen minimal 63.19% (Quintil 3) und maximal 86.90% (Quintil 5). Zudem sind hier lediglich 9 der 25 Zellen in der Transitionsmatrix unbesetzt.

Modellselektion. In den Tabellen 7.11 bis 7.13 wurden Modelle paarweise verglichen, die sich ausschließlich hinsichtlich der berücksichtigten Kovariaten unterscheiden. Die Parametrisierung der Modelle war jeweils konstant. Der Fokus der Analyse lag somit auf der Sensitivität der adjustierten Effektschätzungen infolge der Modifikation der Kovariatenselektion. Nachfolgend sollen schließlich auch die verschiedenen Parametrisierungen einer vergleichenden Betrachtung unterzogen werden, um die Sensitivität des Quintil-Rankings der adjustierten Effektschätzungen infolge der Modifikation der Modellselektion beurteilen zu können.

Tabelle 7.14 zeigt die Transitionen infolge des Wechsels der Parametrisierung für sämtliche der drei betrachteten Modellklassen CAM, VAM und CVA. Die obere Matrix der Tabelle zeigt die Transitionen beim Wechsel der Parametrisierung innerhalb der Modellklasse CAM (Modell 1 vs. Modell 8). Betrachtet man die Diagonale, so ist augenfällig, dass der Modellwechsel mit starken Veränderungen des Quintil-Rankings einhergeht – in erster Linie bei den mittleren Quintilen 2, 3 und 4, in denen jeweils 41.38%, 36.81% und 40.00% der Klassen jeweils identisch kategorisiert sind. Am stabilsten sind wiederum die beiden extremen Quintile 1 und 5 mit 77.93% bzw. 71.72%. Die mittlere Matrix zeigt die Transitionen für die VAM (Modell 4 vs. Modell 11). Hier zeigt sich ein insgesamt ähnliches Ergebnismuster wie in der oberen Matrix, jedoch sind die Effektschätzungen tendenziell stabiler. Die Werte der Diagonale, die den Prozentsatz der nach beiden Modellen identisch kategorisierten Klassen indizieren, betragen 78.62% in Quintil 1, 51.03% in Quintil 2, 44.44% in Quintil 3, 47.59% in Quintil 4 und 74.48% in Quintil 5. Schließlich zeigt die untere Matrix die Transitionen für die CVA

Tabelle 7.14: Veränderungen des Quintil-Rankings der adjustierten Effektschätzungen im Fach Mathematik (MK8): Bedingt lineare vs. lineare Parametrisierung

Contextualized Attainment Model (CAM):						
		Modell 8				
		1	2	3	4	5
Modell 1	1	77.93	20.69	0.69	0.69	0
	2	17.93	41.38	29.66	7.59	3.45
	3	4.17	23.61	36.81	31.94	3.47
	4	0	13.79	24.83	40.00	21.38
	5	0	0.69	7.59	20.00	71.72
Value-Added Model (VAM):						
		Modell 11				
Modell 4	1	78.62	17.24	4.14	0	0
	2	17.93	51.03	26.21	4.14	0.69
	3	3.47	26.39	44.44	22.92	2.78
	4	0	5.52	24.83	47.59	22.07
	5	0	0	0	25.52	74.48
Contextual Value-Added Model (CVA):						
		Modell 14				
Modell 7	1	77.24	19.31	3.45	0	0
	2	20.69	48.97	24.83	5.52	0
	3	2.08	23.61	41.67	24.31	8.33
	4	0	8.28	28.28	44.83	18.62
	5	0	0	1.38	25.52	73.10

Anmerkungen. Die Werte in den Zellen der Tabelle sind Prozent der Klassen pro Zeile, deren Effektschätzung im entsprechenden Quintilsabschnitt liegt. Die zeilenweise Summe beträgt somit jeweils 100%. Anzahl der Klassen: $N = 724$. Anzahl der Klassen pro Zeile bzw. pro Spalte: $n_1 = 145$, $n_2 = 145$, $n_3 = 144$, $n_4 = 145$, $n_5 = 145$.

(Modell 7 vs. Modell 14), wobei hier ein zu den VAM vergleichbares Muster resultiert: Am stabilsten sind wiederum die beiden extremen Quintile 1 und 5 mit 77.24% bzw. 73.10%. Die mittleren Quintile 2, 3 und 4 weisen erneut eine geringere Übereinstimmung auf (48.97%, 41.67% und 44.83%). Die etwas stärkere Stabilität der VAM und der CVA gegenüber der Modifikation der Parametrisierung zeigt sich insbesondere bei Betrachtung der Häufigkeit sehr großer Transitionen (d. h. Wechsel aus Quintil 4 bzw. 5 in Quintil 1 bzw. 2): Dabei nimmt die Häufigkeit der Effektschätzungen, die mehr als 2 Quintile überspringen beim VAM und beim CVA – jeweils im Vergleich zum CAM – ab. Beim Vergleich zwischen VAM und CVA ist die Sensitivität gegenüber dem Wechsel der Parametrisierung wiederum tendenziell höher bei den CVA (vgl. Tabelle F.3 in Anhang F).

Transitionsmatrizen im Fach Deutsch

Kovariatenselektion. Schließlich werde ich auch für den Fachbereich Deutsch Modelle mit jeweils identischer Parametrisierung vergleichen, die sich ausschließlich in der Wahl der Kovariaten unterscheiden.

(1) *CAM vs. VAM*:
Welchen Einfluss auf das Quintil-Ranking der Klassen hat die zusätzliche Berücksichtigung des fachspezifischen Vorwissens (DK3 und DK6) im Fach Deutsch? Tabelle 7.15 zeigt die Transitionsmatrizen, die beim Wechsel vom CAM zum VAM resultieren. Die obere Matrix in Tabelle zeigt die Transitionen für Modelle mit bedingt linearer Parametrisierung mit Interaktionen (Modell 1 vs. Modell 4). Bei Betrachtung der Diagonale fällt auf, dass der Modellwechsel auch im Fach Deutsch mit starken Veränderungen des Quintil-Rankings einhergeht. Erneut zeigt sich diese Sensitivität insbesondere bei den mittleren Quintilen 2, 3 und 4, bei denen jeweils 48.57%, 40.71% und 45.71% der Klassen infolge des Modellwechsels in das gleiche Quintil fallen. Am stabilsten sind die beiden extremen Quintile 1 und 5 mit 77.30% bzw. 70.92%. Die untere Matrix hingegen zeigt die Transitionen für die hinsichtlich der Kovariaten entsprechenden Modelle mit linearer Parametrisierung, in denen potenzielle Interaktionen der Kovariaten nicht modelliert sind (Modell 8 vs. Modell 11). Insgesamt zeigt sich hier ein zu dem in der oberen Matrix sehr ähnliches Ergebnismuster. Die Stabilität in den drei mittleren Quintilen 2, 3 und 4 beträgt jeweils 45.71%, 35.71% bzw.

Tabelle 7.15: Veränderungen des Quintil-Rankings der adjustierten Effektschätzungen im Fach Deutsch (DK8) infolge der zusätzlichen Berücksichtigung des fachspezifischen Vorwissens: CAM *versus* VAM

Bedingt lineare Parametrisierung (inkl. Interaktionen):						
		Modell 4				
		1	2	3	4	5
Modell 1	1	77.30	12.77	6.38	2.13	1.42
	2	18.57	48.57	21.43	6.43	5.00
	3	3.57	27.14	40.71	22.14	6.43
	4	0.71	10.71	26.43	45.71	16.43
	5	0	0.71	4.96	23.40	70.92
Lineare Parametrisierung (ohne Interaktionen):						
				Modell 11		
Modell 8	1	71.63	18.44	6.38	2.84	0.71
	2	22.14	45.71	20.71	9.29	2.14
	3	5.71	27.86	35.71	23.57	7.14
	4	0.71	7.86	32.14	44.29	15.00
	5	0	0	4.96	19.86	75.18

Anmerkungen. Die Werte in den Zellen der Tabelle sind Prozent der Klassen pro Zeile, deren Effektschätzung im entsprechenden Quintilsabschnitt liegt. Die zeilenweise Summe beträgt somit jeweils 100%. Anzahl der Klassen: $N = 702$. Anzahl der Klassen pro Zeile bzw. pro Spalte: $n_1 = 141$, $n_2 = 140$, $n_3 = 140$, $n_4 = 140$, $n_5 = 141$.

44.29%. Hingegen sind die Extrempole der Diagonale – Quintil 1 bzw. Quintil 5 mit 71.63% respektive 75.18% – wiederum am stärksten besetzt.

(2) *VAM vs. CVA*:

Tabelle 7.16 zeigt die resultierenden Transitionsmatrizen, die beim Wechsel vom VAM zum CVA resultieren. Dieser Vergleich adressiert die Sensitivität der Effektschätzungen hinsichtlich der Quintil-Verteilung infolge der zusätzlichen Berücksichtigung der leistungsmäßigen Klassenkomposition. In der oberen Transitionsmatrix in Tabelle 7.16 sind die Transitionen bei bedingt linearer Parametrisierung (mit Interaktionen) dargestellt, wobei hier Modell 4 und Modell 7 verglichen werden. Zwar zeigen sich auch hier deutliche Veränderungen des Quintil-

Tabelle 7.16: Veränderungen des Quintil-Rankings der adjustierten Effektschätzungen im Fach Deutsch (DK8) infolge der zusätzlichen Berücksichtigung von Kompositionsmerkmalen: VAM *versus* CVA

Bedingt lineare Parametrisierung (inkl. Interaktionen):						
		Modell 7				
		1	2	3	4	5
Modell 4	1	81.56	14.18	3.55	0.71	0
	2	16.43	61.43	20.71	1.43	0
	3	2.14	20.71	53.57	22.86	0.71
	4	0	2.14	16.43	60.71	20.71
	5	0	1.42	5.67	14.18	78.72
Lineare Parametrisierung (ohne Interaktionen):						
				Modell 14		
Modell 11	1	90.07	9.93	0	0	0
	2	10.00	81.43	8.57	0	0
	3	0	8.57	80.00	11.43	0
	4	0	0	11.43	80.00	8.57
	5	0	0	0	8.51	91.49

Anmerkungen. Die Werte in den Zellen der Tabelle sind Prozent der Klassen pro Zeile, deren Effektschätzung im entsprechenden Quintilsabschnitt liegt. Die zeilenweise Summe beträgt somit jeweils 100%. Anzahl der Klassen: $N = 702$. Anzahl der Klassen pro Zeile bzw. pro Spalte: $n_1 = 141$, $n_2 = 140$, $n_3 = 140$, $n_4 = 140$, $n_5 = 141$.

Rankings der Klassen basierend auf den jeweils resultierenden Effektschätzungen. Jedoch sind diese Veränderungen im Vergleich zum Wechsel vom CAM zum VAM wiederum geringer. Mit anderen Worten: Die Effektschätzungen sind beim Wechsel vom VAM zum CVA stabiler. So betragen die Werte der Diagonale, welche den Prozentsatz der nach beiden Modellen identisch kategorisierten Klassen indizieren, zwischen minimal 53.57% (Quintil 3) und maximal 81.56% (Quintil 1). Wie auch beim vorangegangenen Vergleich (CAM vs. VAM) sind hier die beiden Quintil-Pole am stabilsten (81.56% in Quintil 1 und 78.72% in Quintil 5). Die untere Matrix in Tabelle 7.16 zeigt die Transitionen für die hinsichtlich der Kovariaten entsprechenden Modelle mit linearer Parametrisierung,

Tabelle 7.17: Veränderungen des Quintil-Rankings der adjustierten Effektschätzungen im Fach Deutsch (DK8) bei bedingt linearer Parametrisierung: Bedingte Unabhängigkeit

Bedingt lineare Parametrisierung (inkl. Interaktionen):						
		Modell 4				
		1	2	3	4	5
Modell 3	1	91.49	8.51	0	0	0
	2	8.57	74.29	15.71	1.43	0
	3	0	15.71	71.43	12.86	0
	4	0	1.43	12.86	76.43	9.29
	5	0	0	0	9.22	90.78
Lineare Parametrisierung (ohne Interaktionen):						
				Modell 7		
Modell 6	1	85.11	12.77	1.42	0.71	0
	2	14.29	61.43	20.00	3.57	0.71
	3	0.71	19.29	57.14	21.43	1.43
	4	0	6.43	20.00	59.29	14.29
	5	0	0	1.42	14.89	83.69

Anmerkungen. Die Werte in den Zellen der Tabelle sind Prozent der Klassen pro Zeile, deren Effektschätzung im entsprechenden Quintilsabschnitt liegt. Die zeilenweise Summe beträgt somit jeweils 100%. Anzahl der Klassen: $N = 702$. Anzahl der Klassen pro Zeile bzw. pro Spalte: $n_1 = 141$, $n_2 = 140$, $n_3 = 140$, $n_4 = 140$, $n_5 = 141$.

in denen potenzielle Interaktionen der Kovariaten nicht modelliert sind (Modell 11 vs. Modell 14). Hier zeigt sich ein insgesamt vergleichbares Ergebnismuster wie in der oberen Matrix. So sind bspw. auch hier die beiden Quintil-Pole am stabilsten mit 90.07% in Quintil 1 und 91.49% in Quintil 5.

(3) *Bedingte Unabhängigkeit*:
Führt die zusätzliche Berücksichtigung beider Vorwissensvariablen (DK3 und DK6) – im Vergleich zur Hinzunahme von DK6 allein – zu substanziellen Veränderungen im Quintil-Ranking der Klassen? Zu diesem Zweck vergleichen wir erneut die Ergebnisse aus den Modellen 3 vs. 4 (bedingt lineare Parametrisierung mit Interaktionen), die in der oberen Matrix von Tabelle 7.17 wiedergege-

ben sind. Die Werte der Diagonale, die den Prozentsatz der nach beiden Modellen identisch kategorisierten Klassen indizieren, betragen 91.49% in Quintil 1, 74.29% in Quintil 2, 71.43% in Quintil 3, 76.43% in Quintil 4 und 90.78% in Quintil 5. Des Weiteren sind 10 der 25 Zellen in der Transitionsmatrix unbesetzt (je 0% der Klassen). Eine demgegenüber geringere Stabilität zeigt sich, wenn Klassenkompositionsmerkmale aus den jeweiligen Klassenstufen – zusätzlich zu den restlichen Kovariaten und den Vorwissensvariablen – berücksichtigt werden. Die entsprechenden Ergebnisse finden sich in der unteren Transitionsmatrix von Tabelle 7.17. Hier betragen die Werte der Diagonale zwischen minimal 57.14% (Quintil 3) und maximal 85.11% (Quintil 1). Zudem sind hier lediglich 4 der 25 Zellen in der Transitionsmatrix unbesetzt.

Modellselektion. Nachfolgend werden schließlich auch im Fachbereich Deutsch die verschiedenen Parametrisierungen einer vergleichenden Betrachtung unterzogen, um die Sensitivität des Quintil-Rankings der adjustierten Effektschätzungen infolge der Modifikation der Modellselektion beurteilen zu können.
Tabelle 7.18 zeigt die Transitionen infolge des Wechsels der Parametrisierung für sämtliche der drei betrachteten Modellklassen CAM, VAM und CVA. Die obere Matrix in Tabelle zeigt die Transitionen beim Wechsel der Parametrisierung innerhalb der Modellklasse CAM (Modell 1 vs. Modell 8). Betrachtet man die Diagonale, so ist augenfällig, dass der Modellwechsel mit starken Veränderungen des Quintil-Rankings einhergeht – in erster Linie bei den mittleren Quintilen 2, 3 und 4, in denen jeweils 45.00%, 42.86% und 51.43% der Klassen kategorisiert sind. Am stabilsten sind wiederum die beiden extremen Quintile 1 und 5 mit 71.63% bzw. 80.85%. Die mittlere Matrix zeigt die Transitionen für die VAM (Modell 4 vs. Modell 11). Hier zeigt sich ein insgesamt recht ähnliches, jedoch tendenziell stabileres Ergebnismuster im Vergleich zur oberen Matrix. Die Werte der Diagonale, die den Prozentsatz der nach beiden Modellen identisch kategorisierten Klassen indizieren, betragen 78.01% in Quintil 1, 57.14% in Quintil 2, 51.43% in Quintil 3, 57.86% in Quintil 4 und 82.27% in Quintil 5. Schließlich zeigt die untere Matrix die Transitionen für die CVA (Modell 7 vs. Modell 14), wobei ein zu den VAM vergleichbares Muster resultiert: Am stabilsten sind wiederum die beiden extremen Quintile 1 und 5 mit 73.05% bzw. 68.79%. Die mittleren Quintile 2, 3 und 4 weisen erneut eine geringere Übereinstimmung auf (51.43%, 42.14% und 50.71%). Auch im Fach Deutsch zeigt sich eine tendenziell stärkere Stabilität der VAM und der

CVA gegenüber der Modifikation der Parametrisierung – insbesondere bei Betrachtung der Häufigkeit sehr großer Transitionen (d. h. Wechsel aus Quintil 4 bzw. 5 in Quintil 1 bzw. 2): Dabei nimmt die Häufigkeit der Klassen, die auf Basis der Effektschätzungen mehr als 2 Quintile überspringen beim VAM und beim CVA – jeweils im Vergleich zum CAM – ab. Beim Vergleich zwischen VAM und CVA ist die Sensitivität gegenüber dem Wechsel der Parametrisierung wiederum tendenziell höher bei den CVA (vgl. Tabelle F.6 in Anhang F).

Zusammenfassung: Veränderungen des Quintil-Rankings

Die Häufigkeiten von Veränderungen des Quintil-Rankings der Effektschätzungen beim paarweisen Modellvergleich bestätigen das bisherige Ergebnismuster.

Innerhalb jeder der beiden Parametrisierungsformen zeigt sich dabei folgender Befund: Durch die Hinzunahme des fachspezifischen Vorwissens in das Adjustierungsmodell ergeben sich deutliche Veränderungen des Quintil-Rankings der klassenspezifischen Effektschätzungen (Hypothese 1.1). Diese zeigen sich gleichfalls beim Wechsel von VAM zu CVA (Hypothese 1.2), jedoch sind die Effektschätzungen dabei tendenziell stabiler. Ist also das fachspezifische Vorwissen bereits gemeinsam mit den weiteren Kovariaten im Modell (VAM) enthalten, reagieren die adjustierten klassenspezifischen Effektschätzungen weniger sensitiv gegenüber der zusätzlichen Aufnahme von Klassenkompositionsmerkmalen (d. h. beim Wechsel vom VAM zum CVA). Des Weiteren ergibt der Vergleich zwischen den beiden Parametrisierungsformen: (a) Ersetzt man – gegeben einem konkreten Kovariatenset – die bedingt lineare durch eine sparsamere lineare Parametrisierung, so ist der Einfluss dieses Modellwechsels am deutlichsten für die CAM. Hier finden sich die stärksten Veränderungen des Quintil-Rankings der Effektschätzungen. Dieser Befund stützt Hypothese 2. (b) Zudem zeigt sich hinsichtlich dieses Kriteriums eine geringere Sensitivität (bzw. höhere Stabilität) der Effektschätzungen für die VAM und auch die CVA infolge des Wechsels der Parametrisierung. Beim Vergleich zwischen VAM und CVA ist die Sensitivität gegenüber dem Wechsel der Parametrisierung wiederum tendenziell höher bei den CVA. Demnach sprechen die Befunde auch hier gegen Hypothese 3. Schließlich stützen auch diese Befunde die Plausibilität von Hypothese 4, da sich das Ergebnismuster nicht in Abhängigkeit vom betrachteten Fachbereich – Mathematik und Deutsch – unterscheidet.

Tabelle 7.18: Veränderungen des Quintil-Rankings der adjustierten Effektschätzungen im Fach Deutsch (DK8): Bedingt lineare vs. lineare Parametrisierung

Contextualized Attainment Model (CAM):						
		Modell 8				
		1	2	3	4	5
Modell 1	1	71.63	24.82	1.42	2.13	0
	2	25.00	45.00	25.71	2.86	1.43
	3	3.57	21.43	42.86	30.00	2.14
	4	0	8.57	24.29	51.43	15.71
	5	0	0	5.67	13.48	80.85
Value-Added Model (VAM):						
				Modell 11		
Modell 4	1	78.01	20.57	1.42	0	0
	2	20.00	57.14	20.71	1.43	0.71
	3	2.14	20.71	51.43	23.57	2.14
	4	0	1.43	25.71	57.86	15.00
	5	0	0	0.71	17.02	82.27
Contextual Value-Added Model (CVA):						
				Modell 14		
Modell 7	1	73.05	24.82	0.71	1.42	0
	2	18.57	51.43	25.00	2.86	2.14
	3	7.86	20.71	42.14	16.43	12.86
	4	0.71	2.86	29.29	50.71	16.43
	5	0	0	2.84	28.37	68.79

Anmerkungen. Die Werte in den Zellen der Tabelle sind Prozent der Klassen pro Zeile, deren Effektschätzung im entsprechenden Quintilsabschnitt liegt. Die zeilenweise Summe beträgt somit jeweils 100%. Anzahl der Klassen: $N = 702$. Anzahl der Klassen pro Zeile bzw. pro Spalte: $n_1 = 141$, $n_2 = 140$, $n_3 = 140$, $n_4 = 140$, $n_5 = 141$.

7.4 Zusammenfassung

Die Ergebnisse der deskriptiven Analyse der Daten und der Analyse der Struktur fehlender Werte wurden dargestellt. Der Fokus des vorliegenden Kapitels lag auf den Ergebnissen des empirischen Modellvergleichs. Die dafür verwendeten Adjustierungsmodelle unterschieden sich hinsichtlich der Kovariatenselektion sowie der Parametrisierung (Modellselektion). Ausgangspunkt bildete das Adjustierungsmodell des Projektes *Kompetenztest.de*, welches im Rahmen der Auswertung und Rückmeldung der Testergebnisse aus den Thüringer Vergleichsarbeiten angewendet wird.

Zur Beurteilung der Sensitivität der adjustierten klassenspezifischen Effektschätzungen gegenüber den Modifikationen der Kovariaten- und Modellselektion wurden fünf Kriterien herangezogen: (a) Caterpillar-Plots, (b) Determinationskoeffizienten $R^2_{Y|\boldsymbol{Z}}$, (c) Korrelationen, (d) Change-Plots und schließlich (e) Transitionsmatrizen. Da Testergebnisse aus Vergleichsarbeiten auf Klassenebene ausgewertet und zurückgemeldet werden, war im Rahmen der vorliegenden Analyse insbesondere der Vergleich der Effektschätzungen einzelner Klassen aus verschiedenen Adjustierungsmodellen relevant. Hierzu eigneten sich insbesondere Change-Plots und Transitionsmatrizen, welche die klassenspezifischen Veränderungen der adjustierten Effektschätzungen abbildeten.

Die Ergebnisse des empirischen Modellvergleichs deuteten insgesamt auf eine hohe Sensitivität der adjustierten klassenspezifischen Effekte hin: Hinsichtlich jedes der gewählten Kriterien zeigten sich Unterschiede in Abhängigkeit von der Wahl der Kovariaten und der Wahl der Parametrisierung. Das Ausmaß und die Richtung des adjustierten Effekts einer Klasse war somit in deutlichem Maße abhängig von der Wahl der Kovariaten bzw. der gewählten Parametrisierung. Dieses Ergebnismuster fand sich zudem in beiden Fachbereichen Mathematik und Deutsch. Die deutlichsten Veränderungen zeigten sich infolge der zusätzlichen Berücksichtigung des fachspezifischen Vorwissens – insbesondere aus Klassenstufe 6 –, also beim Wechsel vom CAM zum VAM. Wurde darüber hinaus auch die leistungsmäßige Klassenkomposition in das Adjustierungsmodell aufgenommen, gab es gleichfalls Veränderungen, die jedoch weniger stark waren. Auch eine sparsamere Parametrisierung führte zu deutlichen Veränderungen in den adjustierten Effektschätzungen. Diese Unterschiede waren geringer zwischen den VAM, d. h., falls das fachspezifische Vorwissen im Adjustierungsmodell enthalten war.

Alles Wissen und alles Vermehren unseres Wissens endet nicht mit einem Schlußpunkt, sondern mit einem Fragezeichen.

HERMANN HESSE (1877 – 1962)

8 Diskussion

Zum Abschluss werde ich die Ergebnisse dieser Arbeit zusammenfassend diskutieren. Es folgt zunächst eine Zusammenfassung des theoretischen Teils sowie der daraus abgeleiteten Implikationen (Abschnitt 8.1). Anschließend werden in Abschnitt 8.2 die Ergebnisse des empirischen Teils dieser Arbeit zusammengefasst. Die zentralen Befunde des empirischen Modellvergleichs werden zudem einer kritischen Diskussion unterzogen. Es folgt eine Reflexion der Grenzen und Generalisierbarkeit der Befunde sowie des sich daraus ergebenden weiteren Forschungsbedarfs in Abschnitt 8.3. Die allgemeine Diskussion wird durch einen Ausblick sowie eine Conclusio in Abschnitt 8.4 abgeschlossen.

8.1 Faire Vergleiche und kausale Effekte

Vergleichsarbeiten sind zu einem etablierten Werkzeug der empirischen Qualitätsentwicklung und -sicherung im Bildungswesen avanciert. In Kapitel 2 habe ich die bildungspolitischen Hintergründe dieser Entwicklung skizziert. Dabei wurde aufgezeigt, dass Vergleichsarbeiten verschiedene, teils nicht-komplementäre Ziele verfolgen. Nicht zuletzt deshalb hat die KMK im Jahr 2012 eine Vereinbarung getroffen, dass die zentrale Funktion in der Unterrichts- und Schulentwicklung der einzelnen Schulen liegen soll. Damit die Ergebnisse dieser standardisierten Testverfahren Ausgangspunkt von Unterrichtsentwicklung sein können, enthalten die Ergebnisrückmeldungen häufig sog. *faire Vergleiche*. Diese fairen Vergleiche standen im Zentrum der vorliegenden Arbeit.

Es besteht ein allgemeiner Konsens, dass bei der Berechnung fairer Vergleiche statistische Adjustierungsverfahren angewendet werden müssen: Für faire Vergleiche müssen Kovariaten – außerschulische Einflussfaktoren auf das Lernen – in der statistischen Auswertung der Testergebnisse berücksichtigt werden. Im Ergebnis sollen diese Vergleiche Aussagen über die Wirksamkeit schulischer Arbeit zulassen. Somit zielen faire

Vergleiche implizit auf die Quantifizierung *kausaler Effekte* des Unterrichts.

Um die Bedingungen und die Grenzen von kausalen Inferenzen explizit zu machen, ist ein theoretisches Fundament unverzichtbar. Aus diesem Grund wurde in Kapitel 3 die Problematik fairer Vergleiche vor dem Hintergrund einer allgemeinen Theorie kausaler Effekte (Steyer et al., 2011) betrachtet, um zu prüfen, ob derartige Vergleiche tatsächlich als kausale Unterrichtseffekte interpretierbar sind. In der allgemeinen stochastischen Theorie kausaler Effekte werden die zentralen Begriffe, die in empirischen Anwendungen und Fragestellungen von Interesse sein können, definiert. Ausgehend von der Definition der theoretischen Größen wurde untersucht, ob die in Vergleichsarbeiten berechneten fairen Vergleiche mit den theoretischen Größen übereinstimmen. Im Ergebnis dieser zunächst analytischen Vorgehensweise zeigten sich drei zentrale Befunde:

(1) Erstens wurde die theoretische Zielgröße definiert: Der Lehrer einer Klasse zielt in aller Regel auf die Optimierung der Unterrichtseffekte bezüglich der jeweils *eigenen* Schülerschaft und nicht etwa bezüglich Schüler im Allgemeinen, die zu großen Teilen gar nicht als Schüler seiner Klasse in Frage kommen. Im Rahmen von Vergleichsarbeiten ist folglich nicht der durchschnittliche kausale Effekt $ACE_{xx'}$ in der Gesamtpopulation von inhaltlichem Interesse, sondern vielmehr der *ACE on the treated*. Ziel ist somit die Schätzung des bedingten kausalen Effekts $CCE_{x;X=x}$ des Unterrichts einer Klasse x bezogen auf die Schülerschaft dieser Klasse x.

(2) Zweitens konnten anhand der Eigenschaften des auf diese Weise definierten Effekts die Möglichkeiten und Grenzen der inhaltlichen Interpretierbarkeit aufgezeigt werden. Diese inhaltlichen Restriktionen hinsichtlich der Interpretierbarkeit sind unabhängig von der Fairness-Problematik. Das bedeutet, dass – selbst wenn wir in der Lage wären, kausale Effekte des Unterrichts auf die Testleistung der Schüler abzubilden – die folgenden zwei Implikationen für die Effektschätzungen Gültigkeit besitzen: (a) Wir betrachten niemals absolute Unterrichtseffekte, sondern stets relative bzw. normative Vergleiche. Insbesondere handelt es sich bei der im Kontext von Vergleichsarbeiten verwendeten Referenz um die soziale Bezugsnorm. Eine darauf basierende Beurteilung der Effektivität des Unterrichts wird spätestens dann problematisch, wenn es keinerlei Varianz in den Lehrerleistungen gibt: Wäre bspw. der Unterricht in allen Klassen gleichermaßen (positiv) wirk-

sam, so wären dennoch sämtliche Effekte (fairen Vergleiche) null. (b) Ein Ranking hinsichtlich des Ausmaßes dieses kausalen Effekts – des *ACE on the treated* – ist nicht ohne weitere Annahmen möglich: Die auf die dargestellte Weise definierten klassenspezifischen kausalen Effekte, d. h. die (X=x)-bedingten kausalen Effekte $CCE_{x;\,X=x}$ des Unterrichts einer Klasse x, sind *zwischen* den Treatment-Bedingungen X=x und X=x' nicht ohne Weiteres miteinander vergleichbar. Ein Ranking dieser Effekte ist daher zunächst nicht bedeutsam.

(3) Und schließlich drittens wurde gezeigt, dass die in Vergleichsarbeiten verwendeten Effektschätzungen nur unter äußerst starken Annahmen den intendierten kausaltheoretischen Größen entsprechen. Diese Annahmen sind im vorliegenden Anwendungsbereich wenig plausibel. Im Kontext von Vergleichsarbeiten kann somit allenfalls von *faireren Vergleichen* ausgegangen werden.

Die bei Vergleichsarbeiten – bzw. bei Schulleistungsuntersuchungen im Kontext von Educational-Accountability-Systemen im Allgemeinen – berechneten adjustierten Maße können folglich nicht als ursächliche Effekte des Unterrichts verstanden werden. Briggs (2008) und Rubin et al. (2004) geben die Empfehlung, diese lediglich als *deskriptive Maße* zu interpretieren. In diesem Sinne können die berechneten adjustierten Effektmaße maximal eine Annäherung an Unterrichtseffekte – und somit lediglich *fairere* Vergleiche – darstellen. Trotz der dargelegten Einschränkungen bergen solche adjustierten Vergleichswerte ein großes Potenzial, um als Informationsbasis für die beteiligten Lehrkräfte zu dienen und Impulse für Unterrichtsentwicklungsmaßnahmen zu geben. So können die Rückmeldungen der Testergebnisse aus Vergleichsarbeiten – und im Besonderen auch die faireren Vergleiche – von den Lehrkräften bspw. dahingehend genutzt werden, das eigene pädagogische Handeln zu reflektieren oder in kollegialen Diskurs über mögliche Ursachen zu treten. Dabei müssen den Adressaten jedoch auch die Grenzen der Interpretation solcher Testergebnisse transparent sein.

8.2 Faire(re) Vergleiche und statistische Adjustierungsmodelle

Die Analyse der Testergebnisse ist eine zentrale Schnittstelle zwischen der Messung – d. h. der Erfassung von Schülerleistungen sowie von außerschulischen Einflussgrößen

des Lernens – und der Ergebnisrezeption seitens der Akteure im Bildungssystem. Alle drei (Messung, Datenanalyse, Rezeption) sind gleichermaßen wichtig (vgl. Abschnitt 2.5.2). Im Zentrum der vorliegenden Arbeit stand jedoch die statistische Auswertung der Testergebnisse aus Vergleichsarbeiten. Diesbezüglich sind faire, kausal interpretierbare Vergleiche zwar theoretisch möglich, im Kontext von Vergleichsarbeiten praktisch jedoch nicht realisierbar. Realistisch hingegen sind fairere Vergleiche, die als deskriptive Maße – im Kontext von Low-Stakes Assessment Systemen – informativen Nutzen haben.

8.2.1 Systematik statistischer Adjustierungsverfahren

Wie kommt man nun zu faireren Vergleichen? Welche Adjustierungsverfahren werden im Kontext von Vergleichsarbeiten de facto angewendet, um fairere Vergleiche zu generieren und in den klassenspezifischen Ergebnisberichten zurückzumelden?

Diesen Fragen widmete sich Kapitel 4. Darin erfolgte zunächst eine Systematisierung der verschiedenen Adjustierungsverfahren bzw. -modelle, die im Rahmen der Ergebnisrückmeldung von Vergleichsarbeiten Anwendung finden. Das Ergebnis dieser Systematisierung wurde in Abschnitt 4.1 dargestellt. Hierbei zeigte sich ein uneinheitliches Bild: In den einzelnen Bundesländern werden unterschiedliche Adjustierungsverfahren angewendet, wobei sich im Wesentlichen vier verschiedene Vorgehensweisen (Strategie I bis IV) differenzieren lassen. Unterschiede bestehen einerseits hinsichtlich der Art und Anzahl der in der Auswertung berücksichtigten Kovariaten (Kovariatenselektion) und andererseits in der Wahl der methodischen Herangehensweise (Modellselektion). So werden bspw. teilweise noch unadjustierte Vergleichswerte zurückgemeldet. Werden Adjustierungen durchgeführt, dann beziehen sich diese auf die Berechnung des (potenziell faireren) Vergleichswertes. Die Differenz eines beobachteten Klassenmittelwertes vom jeweils berechneten adjustierten Vergleichswert soll dann als Maß der Effektivität des Unterrichts interpretiert werden können.

Die verschiedenen Adjustierungsstrategien wurden zudem im Hinblick auf drei Kriterien charakterisiert (Abschnitt 4.1.3), die in engem Zusammenhang stehen und somit eine Triade zur Beurteilung von Adjustierungsverfahren bilden. Die drei Kriterien sind (a) die Fairness, (b) die Testökonomie sowie (c) die Komplexität des im Rahmen der Datenanalyse verwendeten statistischen Modells. Die beiden zuletzt genannten – Testökonomie und Modellkomplexität – adressieren insbesondere die Praktikabilität des

eingesetzten statistischen Verfahrens. Aus methodischer Sicht ist dabei zunächst der Fairness-Aspekt zentral: Nur wenn alle relevanten Kovariaten in der Analyse adäquat, d. h. mittels des richtigen statistischen Modells, berücksichtigt werden, können die berechneten Differenzwerte als ursächliche Effekte des Unterrichts interpretiert werden. In der Praxis der Anwendung von Vergleichsarbeiten – und dies trifft nicht zuletzt auch auf die Praxis der empirischen Bildungsforschung im Allgemeinen zu – spielen jedoch stets auch Praktikabilitätsaspekte wie Testökonomie oder Modellkomplexität eine nicht zu vernachlässigende Rolle. So können bspw. allein aus testökonomischen sowie datenschutzrechtlichen Gründen nicht sämtliche relevante Kovariaten erfasst werden. Und auch das verwendete statistische Modell muss einem Sparsamkeitskriterium genügen. Bei letzterem Argument wird häufig die Verständlichkeit bzw. Kommunizierbarkeit gegenüber den Akteuren im Bildungssystem angeführt. Zusätzlich gibt es jedoch auch statistische und computertechnische Argumente – wie bspw. die Begrenzung der Arbeitsspeicherkapazität eines Rechners – die gleichfalls die Notwendigkeit einer sparsameren Modellierung im Sinne des Parsimonitätsprinzips unterstreichen.

In diesem Sinne lässt sich die Fairness von Leistungsvergleichen aus Vergleichsarbeiten als dimensionale Eigenschaft – und nicht etwa als dichotomes Merkmal – konzeptionalisieren:

(1) *Unfaire Vergleiche*: Werden die außerschulischen Einflussfaktoren und Bedingungen des Lernens, auf die der Lehrer einer Klasse keinen Einfluss hat, in der Auswertung gänzlich außer Betracht gelassen, so sind Leistungsvergleiche als unfair zu erachten. Zudem sind derartige Ergebnisse wenig informativ, da bestehende Unterschiede auf eine Vielzahl potenzieller Ursachen attribuierbar sind.

(2) *Faire Vergleiche*: Erst durch die Berücksichtigung von Kovariaten – also außerschulischer Einflussgrößen des Lernens – sind potenziell faire Vergleiche möglich. Eine notwendige, wenngleich auch nicht hinreichende Bedingung fairer Vergleiche im Sinne der kausalen Definition ist, dass *alle* relevanten Kovariaten in die Analyse einbezogen werden müssen. Dies ist im schulischen Kontext in der Regel nicht realisierbar. Ein fairer, kausal interpretierbarer Vergleich ist folglich aus praktischen Gründen im vorliegenden Anwendungskontext unrealistisch.

(3) *Fairere Vergleiche*: Ein realistisches Ziel hingegen ist die Identifikation und Berücksichtigung der für schulische Leistungen zentralen Kovariaten. Fairere Ver-

gleiche zeichnen sich dadurch aus, dass die in diesem Sinne relevanten Kovariaten in der Auswertung der Testergebnisse mittels eines adäquaten statistischen Modells berücksichtigt werden. Resultierende Unterschiede zwischen beobachteten Werten und adjustierten Referenzwerten lassen sich dann nicht mehr auf Einflüsse dieser Variablen attribuieren.

Des Weiteren wurden die im Rahmen deutscher Vergleichsarbeiten verwendeten Adjustierungsstrategien in den internationalen Kontext eingeordnet. Zu diesem Zweck wurden die Educational-Accountability-Systeme in den Ländern USA und England zum Vergleich herangezogen. In Abschnitt 4.2 wurde herausgearbeitet, dass sich die Adjustierungsstrategien bei Vergleichsarbeiten den Vorgehensweisen in den beiden Ländern USA und England zuordnen lassen. Strategie IVa lässt sich somit den Contextualized Attainment Modellen (CAM) zuordnen, in denen zwar Schüler- und Kompositionsmerkmale als Kovariaten berücksichtigt werden, jedoch explizit nicht das Vorwissen. In Value-Added Modellen (VAM) hingegen ist die Berücksichtigung des Vorwissens (bzw. des Prätests) definitorischer Bestandteil. Auch das Vorgehen des Projektes *Kompetenztest.de* gemäß Strategie IVb lässt sich im Rahmen der Modellierung von VAM verorten. Contextual Value-Added Modelle (CVA) unterscheiden sich davon, indem diese – zusätzlich zu den Schülermerkmalen und deren Vorwissen – auch die leistungsmäßige Klassenkomposition berücksichtigen. Diese Modellklasse findet sich bisher nicht im Rahmen der Ergebnisauswertung von Testergebnissen aus Vergleichsarbeiten.

Diese Einordnung machte zudem auf potenzielle Erweiterungsmöglichkeiten hinsichtlich der Adjustierung von Testergebnissen aufmerksam. So lassen sich bspw. CAM durch die Hinzunahme des fachspezifischen Vorwissens in VAM überführen. Werden weiterhin Kontextmerkmale aufgenommen, die die leistungsmäßige Klassenkomposition betreffen, so erhält man ein CVA. Diese drei Modellklassen – CAM, VAM und CVA – bildeten gleichsam einen Ansatzpunkt für das Design des empirischen Modellvergleichs anhand von Daten aus den Thüringer Kompetenztests, welcher im Zentrum des empirischen Teils dieser Arbeit stand.

8.2.2 Facetten fairer(er) Vergleiche

In Kapitel 5 wurden zunächst die Ergebnisse des theoretischen Teils aggregiert. Ausgehend von einer allgemeinen Theorie kausaler Effekte kann die Missspezifikation eines Adjustierungsmodells im Wesentlichen zwei Ursachen haben: (a) Relevante Variablen

wurden nicht in das Modell aufgenommen (*omitted variables*) oder (b) die modellierte entspricht nicht der wahren funktionalen Form der Abhängigkeit zwischen den im Modell enthaltenen Zufallsvariablen. Somit kann die Problematik fairer(er) Vergleiche im Kontext von Vergleichsarbeiten anhand zweier zentraler Facetten präzisiert werden. Diese betreffen einerseits (a) die Kovariatenselektion, also die Frage, welche Kovariaten im Adjustierungsmodell berücksichtigt werden sollten. Andererseits umfassen diese (b) das Problem der Modellselektion, d. h. die Frage, welches das adäquate statistische Modell zur Schätzung fairerer Vergleiche ist. In der Forschungsliteratur lassen sich drei methodische Zugänge differenzieren, mittels derer diese Teilfragen fairerer Vergleiche analysiert werden können (Abschnitt 5.3). Bisherige Forschungsergebnisse weisen dabei insbesondere auf die Bedeutung des fachspezifischen Vorwissens hin. Die Berücksichtigung des Vorwissens als Kovariate im Adjustierunsgmodell setzt jedoch das Vorliegen längsschnittlicher Daten voraus.

Ausgehend von der Differenzierung der Problematik fairerer Vergleiche in die zwei Facetten Kovariaten- und Modellselektion sowie bisherigen Forschungsbefunden, wurden für die vorliegende Arbeit insbesondere drei zentrale Fragestellungen herausgearbeitet: In Bezug auf die Kovariatenselektion stellt sich erstens die Frage, welchen Einfluss die zusätzliche Berücksichtigung des fachspezifischen Vorwissens sowie der leistungsmäßigen Klassenkomposition auf die klassenspezifischen Effektschätzungen hat. Zweitens wurde der Frage nachgegangen, welchen Einfluss Variationen der Parametrisierung des Adjustierungsmodells (also der Modellspezifikation) auf die klassenspezifischen Effektschätzungen hat. Und schließlich drittens wurde die Frage adressiert, ob es Unterschiede hinsichtlich des Einflusses der Kovariaten- und Modellselektion auf die klassenspezifischen Effektschätzungen zwischen den Unterrichtsfächern Mathematik und Deutsch gibt, in denen Vergleichsarbeiten durchgeführt werden.

Zur Bearbeitung dieser Fragestellungen habe ich im Rahmen der vorliegenden Arbeit den empirischen Zugang gewählt – nicht zuletzt, um die konkreten Gegebenheiten (d. h. die Verteilung und Zusammenhänge der Variablen) bei Daten aus Vergleichsarbeiten bestmöglich abzubilden. Im Rahmen einer empirischen Reanalyse von Daten aus dem Projekt *Kompetenztest.de* wurden unterschiedliche Adjustierungsmodelle auf dieselbe Datenbasis angewendet. Das Ziel dieser Herangehensweise lag darin, durch einen Vergleich der jeweils resultierenden Ergebnisse – d. h. der adjustierten klassenspezifischen Effektschätzungen – Aussagen über die Bedeutung der Variablenselektion sowie der Wahl des statistischen Modells zu ermöglichen. Diese empirische Reanalyse

stellt somit eine Sensitivitätsanalyse dar: Die Sensitivität der Effektschätzungen gegenüber der Modellspezifikation und der Auswahl der Kovariaten wurde analysiert, wobei insbesondere die Bedeutung der Kovariaten Vorwissen (Individualmerkmal) und Leistungsniveau der Schüler einer Klasse (Klassenkompositionsmerkmal) betrachtet wurde.

Das methodische Vorgehen im Rahmen der Sensitivitätsanalyse einschließlich der dafür verwendeten Kriterien sowie das konkrete Design des Modellvergleichs wurde schließlich in Kapitel 6 vorgestellt. Insgesamt 14 verschiedene Modelle wurden angewendet – sowohl für den Fachbereich Mathematik als auch Deutsch. Die einzelnen Modelle lassen sich hinsichtlich der gewählten Kovariaten (Kovariatenselektion) sowie der gewählten Parametrisierung (Modellselektion) unterscheiden.

Da sich die Reanalyse auf Thüringer Kompetenztestdaten bezog, war das Adjustierungsverfahren des Projektes *Kompetenztest.de* (konkret: Strategie IVa) Ausgangspunkt beim Design des Modellvergleichs und diente gleichsam als Referenz des Modellvergleichs. Dieses Modell (Modell 1) kann der Modellklasse der CAM zugeordnet werden. Durch die Hinzunahme des fachspezifischen Vorwissens – dem Kompetenztestergebnis einer früheren Klassenstufe – erhält man ein VAM. Wird zusätzlich auch die leistungsmäßige Klassenkomposition, d. h. der klassenspezifische Mittelwert und die Standardabweichung der Kompetenztestergebnisse früherer Klassenstufen, in das Adjustierungsmodell aufgenommen, so erhält man ein CVA. Die VAM bzw. CVA wurden zudem weiter differenziert in Abhängigkeit davon, ob das klassenspezifische Vorwissen aus Klassenstufe 3 *oder* aus Klassenstufe 6 *oder* sowohl aus Klasse 3 als auch aus Klasse 6 in das Modell aufgenommen wurde. Hinsichtlich der Modellselektion wurde – wiederum ausgehend von der im Projekt *Kompetenztest.de* praktizierten Vorgehensweise – zunächst eine saturierte (CAM) und bedingt lineare (VAM und CVA) Parametrisierung der Adjustierungsmodelle verwendet (Modelle 1 bis 7). Hierbei wurden gleichfalls potenzielle Interaktionen der Kovariaten modelliert. Ferner wurde eine weniger komplexe Parametrisierung für den Modellvergleich gewählt (Modelle 8 bis 14): Sämtliche der in Bezug auf die Kovariatenselektion beschriebenen Modelle – CAM, VAM und CVA – wurden außerdem mit einem linearen Modell ohne Interaktionen modelliert.

8.2.3 Ergebnisse des Modellvergleichs

Der folgende Abschnitt fasst die zentralen Ergebnisse des empirischen Modellvergleichs aus Kapitel 7 in Bezug auf die in dieser Arbeit bearbeiteten Fragestellungen zusammen.

Zudem werden die Befunde einer kritischen Diskussion unterzogen.

Die Ergebnisse des Modellvergleichs weisen insgesamt auf eine deutliche Sensitivität der adjustierten klassenspezifischen Effektschätzungen hin: Hinsichtlich jedes der im Rahmen dieser Arbeit gewählten Sensitivitätskriterien zeigten sich Unterschiede in Abhängigkeit von der Wahl der Kovariaten und der Wahl der Parametrisierung. Im Hinblick auf die in Kapitel 5 postulierten Hypothesen ergibt sich Folgendes:

Hypothese 1: Kovariatenselektion

Die Annahmen bezüglich der Kovariatenselektion wurden in zwei Teilhypothesen formuliert. Dabei bezieht sich Hypothese 1.1 auf das fachspezifische Vorwissen und Hypothese 1.2 auf die leistungsmäßige Klassenkomposition.

Kovariatenselektion: Fachspezifisches Vorwissen. Die zusätzliche Berücksichtigung des fachspezifischen Vorwissens ergänzend zu den weiteren Kovariaten im Adjustierungsmodell führte zu deutlichen Veränderungen der klassenspezifischen Effektschätzungen. Im Vergleich zu den weiteren Modifikationen des Adjustierungsmodells – der Hinzunahme der leistungsmäßigen Klassenkomposition sowie der Vereinfachung der Parametrisierung – war der Einfluss infolge der Hinzunahme des Vorwissens insgesamt am stärksten. So wiesen bspw. die Veränderungen im Determinationskoeffizienten die höchsten Effektstärken auf. Dieses Ergebnismuster zeigte sich in gleicher Weise für jedes der vier anderen Kriterien. Zudem war dieses Ergebnis unabhängig von der Wahl der Parametrisierung des Adjustierungsmodells zu finden, d. h. sowohl innerhalb der bedingt linearen Modelle mit Interaktionen als auch innerhalb der linearen Modelle ohne Interaktionen. Hypothese 1.1 kann somit beibehalten werden. Dies spricht dafür, das fachspezifische Vorwissen in das Adjustierungsmodell einzubeziehen.

Ein detaillierter Blick auf die Ergebnisse weist zudem daraufhin, dass die zusätzliche Berücksichtigung des fachspezifischen Vorwissens aus Klassenstufe 3 – ergänzend zum Vorwissen aus Klassenstufe 6 – zwar einen signifikanten Zuwachs hinsichtlich des Determinationskoeffizienten zur Folge hat, jedoch waren die resultierenden Effektstärken gering. Darüber hinaus zeigte sich auch hinsichtlich der weiteren Kriterien eine vergleichsweise geringe Sensitivität der Effektschätzungen. Mit anderen Worten: Für den faireren Vergleich einer individuellen Klasse macht es einen lediglich marginalen oder keinen Unterschied, ob das Vorwissen aus Klasse 6 und Klasse 3 *oder* nur das

Vorwissen aus Klasse 6 zusätzlich in die Berechnung einfließt[1]. Im Sinne des Parsimonitätsprinzips kann somit tendenziell das sparsamere Modell verwendet werden.

An dieser Stelle muss auf eine Besonderheit im vorliegenden Datensatz hinsichtlich der Struktur fehlender Werte verwiesen werden, die die Belastbarkeit insbesondere der zuletzt genannten Implikation einschränkt. So ist für die fachspezifischen Leistungsvariablen aus Klassenstufe 3 mit jeweils 59% ein sehr hoher Missing-Anteil zu verzeichnen, der vermutlich auf einen Fehler bei der Datenerfassung der Schülerstammdaten im Rahmen des Thüringer Schülerlängsschnitts zurückzuführen ist (vgl. Abschnitt 7.1.2). Dies trifft sowohl auf die Variablen im Fach Mathematik als auch Deutsch zu. Zwar wurden die fehlenden Werte mittels des Verfahrens einer multiplen Imputation ersetzt, Hilfsvariablen im Imputationsmodell verwendet sowie die Plausibilität des Imputationsmodells geprüft (vgl. Abschnitt 7.2). Jedoch setzt dieses Verfahren zum Umgang mit Missings voraus, dass die fehlenden Werte MAR (*missing at random*) sind. Diese Annahme ist – im Gegensatz zur Annahme MCAR (*missing completely at random*) – einer empirischen Prüfung im Sinne einer möglichen Falsifikation nicht zugänglich (vgl. Abschnitt 6.3.2).

Bisher wurde die Frage betrachtet, ob das fachspezifische Vorwissen aus Klassenstufe 3 zusätzlich zum Vorwissen aus Klasse 6 (und den weiteren Kovariaten) berücksichtigt werden sollte. Zwar stellt sich diese Frage nach der Auswahl der Vorwissensvariable in der Regel erst gar nicht, da Vergleichsarbeiten in den meisten Bundesländern ausschließlich in den Klassenstufen 3 und 8 erhoben werden. Demnach liegen zumeist keine Informationen aus standardisierten Leistungstests aus Klasse 6 vor. Jedoch gibt es in diesem Zusammenhang neben den hier dargestellten statistischen Kriterien auch inhaltliche bzw. konzeptionelle Kriterien, die diesbezügliche Aussagen über die Selektion der Kovariaten erlauben. Hierbei spielt die zeitliche Ordnung der Kovariaten eine zentrale Rolle: Wie in Kapitel 3 dargestellt, zeichnen sich Kovariaten insbesondere dadurch aus, dass sie dem Treatment zeitlich vor- oder gleichgeordnet sind. Folglich hat das Treatment per definitionem keinen Einfluss auf die Kovariaten. Wird nun das Testergebnis aus Klassenstufe 6 als Kovariate in das Adjustierungsmodell aufgenommen, so werden gleichfalls potenzielle Effekte der Beschulung *bis* zur sechsten Klassenstufe adjustiert. Dies ist bspw. im Rahmen der empirischen Untersuchung von Effekten von Übergangsentscheidungen – beim Wechsel von der Primar- zur Sekundarstufe – auf die

[1] Dieser Aspekt wurde in Kapitel 7 zum Zwecke der Übersichtlichkeit abgekürzt mit *bedingter Unabhängigkeit* bezeichnet.

Testleistung von Schülern relevant. Will man also explizit den Effekt der Beschulung nach Übergang von der Primar- in die Sekundarstufe untersuchen, so sollten die Testergebnisse aus Klasse 3 – und nicht aus Klassenstufe 6 – als Kovariate verwendet werden. Mit anderen Worten: Will man Unterrichtseffekte innerhalb einer Schulart betrachten – d. h. den Effekt des Unterrichts, der sich auf die Schulzeit nach dem Übergang in die Sekundarstufe attribuieren lässt –, so sollte ausschließlich das Vorwissen aus Klasse 3 berücksichtigt werden. Genau genommen müsste dafür ein Maß des Vorwissens vom Ende der Klassenstufe 4 oder Anfang der Klassenstufe 5 vorliegen. Nimmt man hingegen zusätzlich das fachspezifische Vorwissen aus Klasse 6 hinzu, betrachtet man „nur" den Effekt, den der Unterricht ab Klassenstufe 6 auf die Schülerleistung hat.

Kovariatenselektion: Leistungsmäßige Klassenkomposition. In Kapitel 4 (vgl. Abschnitt 4.2.3) wurde herausgearbeitet, dass für die Evaluation der Wirksamkeit schulischer Arbeit bzw. des Ertrags von Unterricht die Type-B-Effekte – und nicht etwa Type-A-Effekte – von Interesse sind (Raudenbush, 2004; Raudenbush & Willms, 1995; Willms & Raudenbush, 1989). Diese unterscheiden sich von den Type-A-Effekten, indem diese weitere Kovariaten berücksichtigen, die den Schulkontext abbilden. Zu diesem Zweck werden – bspw. im Rahmen von CVA – zusätzlich Kompositionsmerkmale wie bspw. die leistungsmäßige Klassenkomposition berücksichtigt, auf die der Lehrer einer Klasse keinen Einfluss hat. Hypothese 1.2 adressiert diesbezüglich die Bedeutung der leistungsmäßigen Klassenkomposition. Auch die zusätzliche Berücksichtigung der leistungsmäßigen Klassenkomposition – ergänzend zum Vorwissen und den weiteren Kovariaten im Adjustierungsmodell – führte zu Veränderungen der klassenspezifischen Effektschätzungen. Insgesamt ist das Ergebnismuster bezüglich Hypothese 1.2 allerdings weniger deutlich als hinsichtlich Hypothese 1.1 und zudem inkonsistent.

Einerseits hält Hypothese 1.2 über die Bedeutung der leistungsmäßigen Klassenkomposition einer inferenzstatistischen Prüfung nicht stand: Zwar werden die R^2-Differenzen beim Wechsel vom VAM zum CVA bei Anwendung eines sparsameren linearen Adjustierungsmodells statistisch signifikant. Jedoch fällt der Zugewinn hinsichtlich des Kriteriums der Varianzaufklärung infolge der zusätzlichen Berücksichtigung der leistungsmäßigen Klassenkomposition bei bedingt linearer Parametrisierung (inkl. Interaktionen) nicht signifikant aus. Dies spricht gegen die Plausibilität von Hypothese 1.2.

Andererseits zeigte sich insbesondere bei Betrachtung der Ergebnisse auf Ebene einzelner Klassen eine sichtbare Sensitivität der Effektschätzungen. Dies wurde bspw. in

der Darstellung der Veränderung der Effektschätzungen mittels der Change-Plots ersichtlich. Zwar waren hier die Change-Plots im Vergleich zur Hinzunahme des Vorwissens weniger „unruhig“. Und auch die Transitionsmatrizen, die die Veränderungen des Quintil-Rankings der Klassen abbilden, waren vergleichsweise stabiler. Dennoch betrug der Anteil an Inversionen – also der Umkehrung der Richtung des Effekts von positiv zu negativ (bzw. vice versa) – stets ca. 10% aller Klassen.

Die Ergebnisse des inferenzstatistischen R^2-Differenzentests sprechen somit gegen die zusätzliche Berücksichtigung der leistungsmäßigen Klassenkomposition. Jedoch verweist die Sensitivität der klassenspezifischen Effektschätzungen bezüglich der weiteren deskriptiven Kriterien auf Ebene einzelner Klassen darauf, dass auch diese Variablen in das Adjustierungsmodell aufgenommen werden sollten.

Einschränkend sei an dieser Stelle angemerkt, dass die Berücksichtigung dieser Kompositionsmerkmale eine Unterschätzung der Unterrichtseffekte zur Folge haben können: Type-B-Effekte quantifizieren den Ertrag schulischer Praxis, indem für die Eingangsselektivität der Schüler sowie Schulkontextmerkmale (bspw. die leistungsmäßige Komposition der Schülerschaft) kontrolliert wird. Raudenbush und Willms (1995) verstehen die Schulleistung eines Schülers (Outcome) dabei als additive Funktion von vier Faktoren (individuelle Schülermerkmale, Messfehler, Schulkontext und Schulpraxis) und ignorieren somit die potenziell multiplikative Verknüpfung dieser Facetten schulischer Arbeit. Jedoch zeigen empirische Befunde, dass es durchaus Interaktionseffekte zwischen der Schulkomposition und der Schulpraxis geben kann (vgl. Iturre, 2005; Opdenakker & Van Damme, 2007). In diesem Fall führt die Berücksichtigung von Kompositionsmerkmalen der Schule zu einer Unterschätzung der Schuleffekte. Andererseits führt die Vernachlässigung solcher Kompositionsmerkmale potenziell zu einer Überschätzung von Schuleffekten, da die Effekte, die einzig auf die Komposition zurückzuführen sind, nicht von den tatsächlichen Schuleffekten separiert werden (vgl. Timmermans et al., 2011).

Hypothese 2: Modellselektion

Eine sparsamere lineare Parametrisierung führte zu Veränderungen der adjustierten klassenspezifischen Effektschätzungen. Diese Veränderungen zeigten sich hinsichtlich jedes einzelnen der im Rahmen dieser Arbeit gewählten Kriterien, die zur Beurteilung der Sensitivität der Effektschätzungen herangezogen wurden. So zeigten sich bspw.

signifikante Unterschiede zwischen den Determinationskoeffizienten der beiden CAM, die sich ausschließlich hinsichtlich der gewählten Parametrisierung unterschieden. Und auch auf Ebene einzelner Klassen zeigten sich hier deutliche Veränderungen: Der Anteil an Inversionen infolge des Wechsels der Parametrisierung lag bei 18% im Fach Mathematik und 15% im Fach Deutsch. Folglich wird Hypothese 2 bezüglich des Haupteffektes der Parametrisierung (Modellselektion) auf die klassenspezifischen Effektschätzungen zunächst beibehalten. Die Annahme eines bedingt linearen Zusammenhangs, inklusive der Modellierung von Interaktionen, ist demnach einer sparsameren linearen Parametrisierung (ohne Interaktionsterme) vorzuziehen. Ob es hinsichtlich des Einflusses der Parametrisierung Unterschiede in Abhängigkeit von der Kovariatenselektion gibt, wird in der folgenden Hypothese thematisiert.

Hypothese 3: Kovariatenselektion vs. Modellselektion

Die dritte Hypothese betrifft die Interaktion zwischen Kovariaten- und Modellselektion. Diese adressiert somit die Frage, ob es differentielle Effekte der Parametrisierung auf die adjustierten klassenspezifischen Effektschätzungen in Abhängigkeit von der Kovariatenselektion gibt. Hier habe ich basierend auf bisherigen Forschungsbefunden (vgl. Abschnitt 5.3) angenommen, dass die Selektion relevanter Kovariaten einen größeren Einfluss auf die Effektschätzungen hat als die Parametrisierung (Modellselektion). Demnach sollte der Einfluss der Parametrisierung auf die Effektschätzungen geringer sein, je mehr relevante Variablen im Adjustierungsmodell des Projektes *Kompetenztest.de* enthalten sind. Auch bezüglich dieser Hypothese ist das Ergebnismuster weniger deutlich als hinsichtlich Hypothese 1.1 über die Bedeutung des fachspezifischen Vorwissens und zudem inkonsistent.

Zwar hält diese dritte Hypothese einer inferenzstatistischen Prüfung hinsichtlich der R^2-Differenzen stand: So wurden die Veränderungen im Determinationskoeffizienten zwischen Modellen mit komplexerer Parametrisierung im Vergleich zur linearen Parametrisierung für die CAM zunächst signifikant. War hingegen das fachspezifische Vorwissen MK3 und MK6 im Fachbereich Mathematik bzw. DK6 im Fachbereich Deutsch zusätzlich im Adjustierungsmodell enthalten, waren die Unterschiede im Determinationskoeffizienten zwischen Modellen mit linearer vs. mit bedingt linearer Parametrisierung nicht mehr signifikant. Und auch zwischen den CVA wurden die R^2-Differenzen nicht signifikant. Die Ergebnisse des inferenzstatistischen R^2-Differenzentests sprechen

somit für ein sparsameres lineares Modell, wenn das fachspezifische Vorwissen zusätzlich im Modell enthalten ist.

Jedoch sei hier einschränkend angemerkt, dass der R^2-Differenzentest wesentlich von der Modellkomplexität der zu vergleichenden Modelle sowie der damit verbundenen Differenz in der Anzahl zu schätzender Parameter einhergeht. Bei den 14 betrachteten Modellen umfassen insbesondere die Modelle mit bedingt linearer Parametrisierung (inkl. Interaktionen) eine sehr große Parameterzahl. Und auch die Differenz der Parameteranzahl ist teilweise sehr groß: Die Modelle mit identischem Kovariatenset aber ungleicher Parametrisierung unterscheiden sich um einige hundert oder sogar einige tausend Parameter. Dies wird in den Tabellen 7.7 und 7.10 ersichtlich, denn die Zählerfreiheitsgrade df_1 geben hier jeweils die Differenz der Parameteranzahl jeweils genesteter Modelle wieder. Die minimale Differenz liegt zwischen den CAM mit $df_1 = 313$ Parametern vor. Die maximale Differenz zwischen den CVA beträgt $df_1 = 5\,109$. Je größer die Differenz der Parameteranzahl – und folglich je größer die Zählerfreiheitsgrade –, desto kleiner ist die Teststatistik (der F-Wert). Daraus resultiert ein teilweise widersprüchliches Ergebnismuster zwischen den inferenzstatistischen Kriterien und den weiteren, mehr deskriptiven Kriterien. So zeigte sich bereits hinsichtlich der Effektstärken ein der Hypothese 3 tendenziell widersprechendes Ergebnismuster: Zwar sank die R^2-Differenz zwischen Modellen unterschiedlicher Parametrisierung beim Wechsel vom CAM zum VAM zunächst ab, jedoch stieg diese beim Wechsel vom VAM zum CVA wieder an. Auch bei den anderen Kriterien – Korrelationen, Change-Plots und Transitionsmatrizen – zeigte sich dieses Ergebnismuster. Diese deskriptive Befundlage spricht somit gegen die sparsamere lineare Parametrisierung, bei der potenzielle Interaktionen zwischen den Kovariaten nicht modelliert werden.

Diese widersprüchliche Befundlage ist für sich genommen nicht problematisch – können doch höhere Determinationskoeffizienten Ausdruck einer Überanpassung (engl.: *overfitting*) des Regressionsmodells sein. Die Korrelationen der Effektschätzungen auf Klassenebene, die Change-Plots und die Transitionsmatrizen adressieren zudem insbesondere die Frage nach der Stabilität der klassenspezifischen Effektschätzungen. Diese beiden Kriterien – das Stabilitätskriterium und das Kriterium der Varianzaufklärung – sind zwar nicht unabhängig voneinander, sie können jedoch nicht per se gleichgesetzt werden. So können die adjustierten klassenspezifischen Effektschätzungen einzelner Klassen deutlich von der Parametrisierung (Modellselektion) abhängen, selbst wenn die Linearitätshypothese basierend auf den Ergebnissen des R^2-Differenzentests bei-

behalten wird. Sollen die adjustierten Effektschätzungen tatsächlich für diagnostische Zwecke auf Klassenebene verwendet werden, sind gegebenenfalls komplexere Modelle vorzuziehen – auch wenn Signifikanztests hinsichtlich der Varianzaufklärung ein sparsameres Modell bevorzugen.

In diesem Zusammenhang wird eine weitere potenzielle Einschränkung relevant, die das in Abschnitt 4.1.3 erläuterte Praktikabilitätskriterium der Komplexität des statistischen Modells betrifft: Mit einem hinsichtlich der Parameteranzahl komplexeren Modell – wie dem im Rahmen der vorliegenden Anwendung verwendeten bedingt linearen Adjustierungsmodell – steigen die Anforderungen (1) einerseits an die computertechnischen Ressourcen und (2) andererseits an die zur Verfügung stehenden Daten. Beide Probleme traten im Rahmen der Reanalyse des vorliegenden Datensatzes – speziell bei den Modellen mit bedingt linearer Parametrisierung – auf:

(1) Das Problem begrenzter computertechnischer Ressourcen trat bereits im Rahmen der Berechnung der VAM mit bedingt linearer Parametrisierung auf. Erst die Verwendung leistungsstarker Rechner ermöglichte die Parameterschätzung der VAM und CVA.

(2) Die zweite Herausforderung bestand hinsichtlich der zur Verfügung stehenden Daten: So musste das komplexeste Modell (Modell 7) um zwei Prädiktorvariablen reduziert werden, da das Modell nicht identifiziert war. Das vollständige Modell umfasste 20 480 Parameter. Jedoch umfasste der vorliegende Datensatz lediglich $N = 12\,708$ Beobachtungen. Es resultierte eine $12\,708 \times 20\,480$-Designmatrix. Folglich waren die Parameter nicht schätzbar. In dem reduzierten Modell wurden die Standardabweichungen der beiden Vortestvariablen als Prädiktoren ausgeschlossen[2]. In dem auf diese Weise vereinfachten Modell konnte die Anzahl der Parameter auf 5 120 reduziert werden. Dieses Modell konnte schließlich mittels der freien Software R berechnet werden.

Aufgrund der Ergebnisse dieser Arbeit wird geschlussfolgert, dass die Wahl der richtigen Kovariaten von vorrangiger Bedeutung ist. Die korrekte Modellwahl (Parametrisierung) in empirischen Anwendungen unterliegt – gegeben dem limitierten Umfang an

[2]Derartige Probleme traten folglich nicht im Rahmen der Berechnung der sparsameren, linearen Adjustierungsmodelle auf. Um potenzielle Unterschiede in den adjustierten Effektschätzungen zwischen den Modellen 7 und 14 auf die Modifikation der Parametrisierung attribuieren zu können, wurden auch in Modell 14 die Standardabweichungen der beiden Vortestvariablen als Prädiktoren ausgeschlossen.

Daten – stets natürlichen Beschränkungen. Die empirischen Befunde sprechen insgesamt für die Anwendung komplexerer Modelle, auch wenn inferenzstatistische Kriterien wie der R^2-Differenzentest ab einem bestimmten Kovariatenset sparsamere Modelle präferieren.

Hypothese 4: Generalisierung über Fächer

Die in den ersten drei Hypothesen postulierten Zusammenhänge – d. h. die Sensitivität der adjustierten klassenspezifischen Effektschätzungen gegenüber der Kovariaten- und Modellselektion – lassen sich sowohl im Fach Mathematik als auch im Fach Deutsch finden. Dabei ist das Ergebnismuster in beiden Fächern konkordant und tendenziell etwas deutlicher für den Fachbereich Deutsch. Somit wird auch Hypothese 4 nicht falsifiziert, da sich die Ergebnismuster infolge der Modifikation der Adjustierungsmodelle nicht fachspezifisch zeigten. Die dargestellten Implikationen hinsichtlich der Modifikation des Adjustierungsmodells lassen sich somit sowohl für Kompetenztestergebnisse im Fachbereich Mathematik als auch Deutsch ziehen.

8.3 Grenzen und kritische Reflexion

Die empirische Reanalyse der Thüringer Kompetenztestdaten lieferte empirische Evidenz über die Bedeutung der Kovariatenselektion und der Parametrisierung des Adjustierungsmodells im Kontext fairerer Vergleiche in Schulleistungsuntersuchungen – speziell in den landesweiten Vergleichsarbeiten. Dabei bestätigte sich insbesondere die Bedeutung der Kovariatenselektion – konkret des fachspezifischen Vorwissens sowie der leistungsmäßigen Klassenkomposition. Dies zeigte sich unabhängig vom betrachteten Fachbereich. Zudem stellt die verwendete Vorgehensweise ein evidenzbasiertes Verfahren dar, welches zur Entwicklung neuer oder Weiterentwicklung bestehender Adjustierungsverfahren bei der Ergebnisauswertung und -rückmeldung von Testleistungen aus Vergleichsarbeiten – also zur Erstellung fairerer Vergleiche – beitragen kann. Jedoch lassen sich die Ergebnisse der vorliegenden Untersuchung nicht ohne Weiteres verallgemeinern. Die Grenzen der vorliegenden Untersuchung – u. a. in Bezug auf die Generalisierbarkeit der Ergebnisse – sowie die sich daraus ergebenden Anforderungen an zukünftige Untersuchungen stehen im Fokus des folgenden Abschnitts.

8.3.1 Weiterer Forschungsbedarf

Im Rahmen der jährlichen Erhebung der Thüringer Kompetenztests werden neben den Leistungsdaten verschiedene Schülermerkmale erfasst, die bei der Berechnung fairerer Vergleiche als Kovariaten dienen. Dabei wird u. a. Anzahl der Bücher im Elternhaus erfasst, die sich auch in anderen Untersuchungen als geeigneter Indikator des sozioökonomischen Status (SES) bzw. der Bildungsnähe erwiesen hat (vgl. Bos et al., 2003; M. D. Evans et al., 2010). Damit sind jedoch zwei Einschränkungen verbunden:

(1) Erstens bildet die Bücherfrage nur eine Facette des sozialen Hintergrundes der Schüler ab. In der soziologischen Forschung werden verschiedene Facetten der sozialen Herkunft unterschieden, die sich auf den sozialen und ökonomischen Status sowie die kulturelle Praxis einer Familie beziehen und als Ressourcen sozialer Entwicklungsmilieus fungieren können. So unterscheidet Bourdieu (1983) zwischen sozialem, kulturellem und ökonomischem Kapital, das in der Forschungspraxis mit jeweils unterschiedlichen Indikatoren operationalisiert wird. Die Anzahl der Bücher ist ein Indikator des kulturellen Kapitals einer Familie. Hohe kulturelle Ressourcen gehen zwar häufig mit hohen ökonomischen Ressourcen einher, müssen es jedoch nicht (vgl. Ehmke & Siegle, 2005). Zukünftige Untersuchungen könnten somit die Frage adressieren, ob für das Adjustierungsvorgehen des Projektes *Kompetenztest.de* weitere bzw. andere Indikatoren der sozialen Herkunft geeigneter sind.

(2) Zweitens wird die Bücherfrage aus datenschutzrechtlichen Gründen seit dem Schuljahr 2008/2009 nur noch in anonymisierter Form auf Klassenebene erhoben. Das bedeutet, dass im vorliegenden Datensatz für jede Klasse lediglich der Mittelwert des SES über die Schüler dieser Klasse vorliegt. Daher haben alle Schüler einer konkreten Klasse die gleiche Ausprägung auf der Variable SES (vgl. Abschnitt 6.2.2). Folglich wird bei der Berechnung fairerer Vergleiche der SES lediglich als Kompositionsmerkmal – und nicht als individuelles Merkmal eines Schülers – berücksichtigt. Dabei besteht in der vorliegenden Untersuchung ein deutlicher Zusammenhang zwischen diesem klassenspezifischen SES und der individuellen Testleistung der Schüler ($r_{\text{Mathematik}} = .55$ und $r_{\text{Deutsch}} = .52$). Die entsprechenden Korrelation auf Klassenebene ist noch deutlicher ($r_{\text{Mathematik}} = .71$ und $r_{\text{Deutsch}} = .73$). Dennoch besteht die Gefahr eines ökologischen Fehlschlus-

ses (engl.: *ecological fallacy*; vgl. Robinson, 1950) über die Bedeutsamkeit dieser Variable: Diese Korrelationen lassen keinen Schluss zu, wie stark der individuelle SES der Schüler mit der individuellen Testleistungsvariable korreliert. Inwiefern diese Tatsache die faireren Vergleiche, d. h. die adjustierten klassenspezifischen Effektschätzungen, beeinflusst, kann im Rahmen der vorliegenden Untersuchung nicht geklärt werden. Dazu bedarf es weiterer Untersuchungen, in denen die entsprechenden Informationen für jeden einzelnen Schüler vorliegen.

Eine weitere Einschränkung hinsichtlich der Belastbarkeit der vorliegenden Befunde wurde bereits in Abschnitt 8.2.3 diskutiert und betrifft die Struktur fehlender Werte. Der empirische Modellvergleich basiert auf einem Datensatz, bei dem auf einzelnen Variablen bis zu 59% fehlende Werte vorlagen. Zwar wurden die fehlenden Werte mittels des Verfahrens einer multiplen Imputation ersetzt. Dennoch basiert die multiple Imputation auf Annahmen, die einer empirischen Prüfung nicht zugänglich sind. Da der hohe Missing-Anteil vermutlich aufgrund eines Fehlers bei der Datenerfassung in dem vorliegenden Erhebungsjahrgang verursacht wurde, sollten zukünftige Untersuchungen nicht mit derart hohen Missing-Raten konfrontiert sein. Für zukünftige Untersuchungen stellt sich somit die Frage der Replizierbarkeit vorliegender Befunde.

8.3.2 Generalisierbarkeit der Ergebnisse

Die vorliegenden Befunde sind zwar im spezifischen Kontext des Adjustierungsverfahrens des Projektes *Kompetenztest.de* gewonnen. Diese sind jedoch auch für andere Bundesländer und damit auch für die weiteren Adjustierungsstrategien relevant. Dabei lassen sich die vorliegenden Befunde nicht ohne Weiteres auf andere Bundesländer und die restlichen Adjustierungsstrategien generalisieren – wenngleich auch keine gänzlich widersprechendes Befundmuster zu erwarten sind.

Generalisierbarkeit auf andere Bundesländer

Bereits in Abschnitt 4.1.3 wurde herausgearbeitet, dass die Wahl eines geeigneten Adjustierungsverfahrens – insbesondere hinsichtlich der Kovariatenselektion – auch von der Verteilung der relevanten Kovariaten in der jeweils betrachteten Population abhängt. So müssen stets auch die spezifischen Gegebenheiten eines Bundeslandes – bezogen auf die Zusammensetzung der Schülerschaft hinsichtlich relevanter Kovariaten – bei der

Wahl der Adjustierungsstrategie besondere Berücksichtigung finden. Ein – gewiss sehr extremes und gleichsam wenig realistisches – Beispiel soll dies verdeutlichen. Dabei gehen wir davon aus, dass sich sowohl theoretisch als auch empirisch herausgestellt hat: Brillenträger profitieren hinsichtlich ihrer Leistung deutlich mehr von einem pädagogischen Treatment als Nicht-Brillenträger. Das Merkmal *Brille-Tragen* hat sich somit als eine zentrale Kovariate der Schulleistung erwiesen. Falls es in Thüringen jedoch ausschließlich Brillenträger gibt, ist die Variable eine Konstante und gleichfalls nicht als Kovariate im Rahmen der bundeslandspezifischen Adjustierung nutzbar. Zudem ist in diesem Fall lediglich die Schätzung des bedingten Effekts für Brillenträger möglich. Dies stellt gleichsam eine zusätzliche Herausforderung im Rahmen der Formulierung allgemein verbindlicher Standards bzw. Richtlinien bei der Erstellung fairerer Vergleiche dar. Man muss also in jedem Bundesland prüfen, welche Kovariaten relevant sind. Zukünftige Untersuchungen sollten dies berücksichtigen.

Generalisierbarkeit auf andere Adjustierungsverfahren

Zudem beziehen sich die vorliegenden Ergebnisse insbesondere auf die in Kapitel 4 vorgestellten Adjustierungsstrategien IVa und IVb, die beide im Rahmen des Projektes *Kompetenztest.de* angewendet werden. Eine Generalisierung der Ergebnisse auf die restlichen Adjustierungsstrategien, die zu faireren Vergleichen führen sollen (Strategie II und III), ist jedoch nicht ohne Weiteres möglich: Die verschiedenen Strategien zeichnen sich nicht nur durch ein jeweils spezifisches Kovariatenset, sondern gleichfalls ein spezifisches statistisches Verfahren aus. Demnach sind die Adjustierungsstrategien nicht unmittelbar miteinander vergleichbar.

So ist bspw. ein direkter Vergleich von Strategie IIIa – der in Nordrhein-Westfalen angewendeten Standorttypisierung – und Strategie IVb des Projektes *Kompetenztest.de* anhand des vorliegenden Datensatzes nicht möglich, da die dafür notwendigen Informationen über die Standorttypen der Thüringer Schulen nicht vorliegen. Jedoch ist ein derartiger Vergleich ebenso wenig anhand von Daten aus Vergleichsarbeiten möglich, die in Nordrhein-Westfalen erhoben wurden. Der Grund hierfür liegt in der querschnittlichen Datenstruktur: Informationen über das fachspezifische Vorwissen, die mittels Testleistungen aus Vergleichsarbeiten vorangegangener Erhebungszeitpunkte erhoben wurden, liegen nicht vor. Zudem lassen sich die Ergebnisse – selbst beim Vorliegen der gleichen Kovariatensets – nicht direkt vergleichen, da die beiden zugrunde liegenden

Modelle keine genestete Struktur aufweisen.

Gleichwohl können komparative Untersuchungen dieser Art zur Weiterentwicklung fairerer Vergleiche beitragen. Ein derartiger Vergleich erlaubt ebenso wenig Aussagen über *das richtige Verfahren* zur Erstellung fairer Vergleiche wie die hier vorgestellte Sensitivitätsanalyse. Der Grund dafür liegt im Fehlen einer kausalen Benchmark (vgl. Abschnitt 5.3.1). Welchen Nutzen hat dann ein solcher Vergleich? Diese komparative Vorgehensweise ermöglicht Aussagen über die Sensitivität bzw. Stabilität der resultierenden Ergebnisse für einzelne Klassen. Dabei sollten die Ergebnisse der verschiedenen Verfahrensweisen möglichst stabil sein. Unterschiede in den klassenspezifischen Ergebnissen hingegen indizieren potenzielle Weiterentwicklungsmöglichkeiten des jeweils sparsameren Verfahrens. Auf diese Weise kann eine solche Vorgehensweise trotz der erwähnten Einschränkungen im Rahmen von Best-Practice-Entscheidungen zur Entwicklung neuer oder Weiterentwicklung bestehender Adjustierungsverfahren bei der Ergebnisauswertung und -rückmeldung von Testleistungen aus Vergleichsarbeiten – zur Erstellung fairerer Vergleiche – genutzt werden.

8.4 Conclusio und Ausblick

Faire, kausal interpretierbare Vergleiche sind zwar theoretisch möglich, jedoch im Kontext von Schulleistungsuntersuchungen wie bspw. Vergleichsarbeiten m. E. praktisch nicht realisierbar. Auf öffentliche Rankings im Rahmen der Rechenschaftslegung der Schulen oder anderen Konsequenzen im Rahmen von High-Stakes Assessment Systemen sollte weiterhin verzichtet werden, denn die im Rahmen dieser Modelle geschätzten Effekte „… that can be presently derived from the information in educational accountability systems are unlikely to be robust enough to support high-stakes causal inferences about school quality in educational settings“ (Briggs & Wiley, 2008, S. 188). Realistisch hingegen sind fairere Vergleiche, die als deskriptive Maße im Kontext von Low-Stakes Assessment Systemen informativen Nutzen haben. Bei der Erstellung solcher faireren Vergleiche sollte – wenn immer möglich – das fachspezifische Vorwissen der Schüler einbezogen werden.

8.4.1 Ausblick

Wie können fairere Vergleich dennoch zur Beantwortung kausaler Fragestellungen dienen? Auch als deskriptive Maße können fairere Vergleiche im Rahmen kausaler Fragestellungen zur Unterrichtsqualität und Unterrichtsentwicklung nutzbar gemacht werden. Zwei mögliche Forschungsansätze dieser Art seien nachfolgend als kurzer Ausblick skizziert.

Evaluation von Educational-Accountability-Systemen. Dies betrifft einerseits die Evaluation der Wirksamkeit von Vergleichsarbeiten und den damit verbundenen klassenspezifischen Rückmeldungen fairerer Vergleiche. Hierbei geht es um die Frage nach den kausalen Effekten von testbasierten Evaluationssystemen im Bildungssystem (vgl. Rubin et al., 2004). Im Rahmen solcher Untersuchungen geht es folglich nicht um die Schätzung des kausalen Effekts von Unterricht und Schule auf die Testleistung der Schüler. Stattdessen wird bspw. der kausale Effekt der Durchführung von Vergleichsarbeiten und der damit verbundenen Rückmeldung fairerer Vergleiche fokussiert.

Ein Beispiel für eine solche Herangehensweise soll dies verdeutlichen: In einer Studie von McCaffrey, Stuart, Rubin und Zanutto (2006) wurde der Effekt des Einsatzes eines Value-Added Assessment Systems untersucht, welches in das Educational-Accountability-Systems des Staates Pennsylvania eingeführt wurde. Zu diesem Zweck wurde die Testleistung von Schülern aus Schulbezirken betrachtet, in denen die klassenspezifischen Rückmeldungen Effektschätzungen aus Value-Added Modellen (VAM) enthielten. Diese Testergebnisse wurden der Testleistung von Schülern in Schulbezirken gegenübergestellt, die eine hinsichtlich leistungsrelevanter Merkmale ähnliche Schülerschaft aufwiesen, in denen jedoch (noch) keine VAM-Schätzungen zurückgemeldet wurden. Zusätzlich wurden qualitative Informationen darüber erhoben, wie die Schulen bzw. Klassen die Ergebnisse aus den VAM-Analysen nutzten.

Im Zuge solcher empirischer Untersuchungen – bspw. im Rahmen systematischer Rezeptionsforschung – können somit kausale Effekt der Implementation von Low-Stakes Assessment Systemen wie den Vergleichsarbeiten inklusive der damit verbundenen Rückmeldesysteme evaluiert werden.

Unterrichtsqualitätsforschung. Die Erhebung und systematische Evaluation von Prozessmerkmalen des Unterrichtsgeschehens, die zu einem gelingenden Unterricht

führen, ist Gegenstand der Unterrichtsqualitätsforschung (z. B. Gräsel & Göbel, 2011). Auch hier können fairere Vergleiche als Ansatzpunkt eines kausalen Forschungsprozesses zur Untersuchung solcher Faktoren gelingenden Unterrichts genutzt werden. So können fairere Vergleiche zur Identifikation von Klassen genutzt werden, die sich durch besonders positive klassenspezifische Effektschätzungen auszeichnen, denn diese lassen sich nicht den außerschulischen Einflussgrößen des Lernens zuschreiben. Sind diese Klassen identifiziert, so könnte ein weiteres Forschungsprogramm oder -projekt die Bedingungen gelingenden Unterrichts untersuchen.

Ein Beispiel hierfür ist die *California Best Practices Study* (CBPS) – einem dreijährigen Forschungsprojekt in Californischen Schulen (Arbeit, Canter, Dabrowski, Dailey & Oberman, 2007; Oberman, 2005). Im Rahmen der CBPS wurden *high performing schools* basierend auf Ergebnissen aus den jährlich durchgeführten standardisierten Schulleistungstests identifiziert. Diese Schulen wurden mit einer (Kontroll-)Gruppe von Schulen verglichen, die hinsichtlich der demographischen Zusammensetzung der Schülerschaft komparabel war, jedoch vergleichsweise schlechte Testergebnisse aufwies. Diese beiden Gruppen wurden mittels qualitativer und quantitativer Methoden hinsichtlich verschiedener Faktoren (z. B. didaktische und curriculare Ziele, Personalauswahl, Führungsverhalten, Nutzung von Daten zur Selbstevaluation, Anpassung des Unterrichts etc.) untersucht, um die Bedingungen gelingenden Unterrichts aufzudecken. Die Identifikation solcher Bedingungen erlaubt darüber hinaus bspw. die Entwicklung spezifischer didaktischer oder pädagogisch-psychologischer Maßnahmen.

Mittels eines solchen Untersuchungsansatzes wird der Fokus weg von der Schätzung kausaler Unterrichts- oder Lehrereffekte und hin zur Schätzung kausaler Effekte einer oder mehrerer spezifischer didaktischer Maßnahmen gelenkt. Bei der Evaluation solcher Maßnahmen sind schließlich auch experimentelle Forschungsdesigns – dem *Goldstandard* kausaler Inferenz – denkbar, in denen mittels (bedingter) Randomisierung die (bedingte) Unverfälschtheit hergestellt werden kann (vgl. Abschnitt 3.4). In diesem Sinne können fairere Vergleiche in Vergleichsarbeiten als Fundament bzw. Startpunkt eines iterativen Forschungsprozesses mit dem Ziel kausaler Inferenz fungieren. „It is in this sense that an emphasis on causal estimation can go long way in moving the basis of causal inferences from speculation to science“ (Briggs & Wiley, 2008, S. 188). Daraus resultierende empirische Befunde unterstreichen den Nutzen einer evidenzbasierten Bildungsforschung bspw. hinsichtlich der Unterrichtsentwicklung.

8.4.2 Mindestanforderungen für Low-Stakes Assessment

Standardisierte und standardbezogene Schulleistungstests – wie die landesweiten Vergleichsarbeiten in Deutschland oder die Key Stage Tests in England – sind wesentliche Komponenten einer evidenzbasierten Steuerung des Bildungssystems, in dessen Rahmen die erreichten Schülerleistungen bzw. -kompetenzen als ein zentrales Kriterium zur Evaluation der Leistungsfähigkeit eines Bildungssystems genutzt werden. So äußerte sich Prof. Dr. Hans Anand Pant, Direktor des IQB, in einem Interview mit der Wochenzeitung *DIE ZEIT* zur Bedeutung von Vergleichsarbeiten: „Externe Leistungstests sind alternativlos. Der Bewertungshorizont eines Lehrers ist immer die eigene Klasse, vielleicht noch die eigene Schule. Deshalb benötigen wir in Abständen Vergleiche, die sich an anerkannten Kompetenzstandards orientieren“ (Spiewak, 2011).

Vergleichsarbeiten sind somit nicht mehr wegzudenken aus einem modernen Educational-Accountability-System, in dem schulische und klassenspezifische Qualitätssicherungsmaßnahmen auf testbasierten, empirischen Evaluationssystemen beruhen. Evaluation ist dabei immer ein komparativer Prozess, denn Evaluation geht stets mit Vergleichen einher. Hier können fairere Vergleiche einen wichtigen und konstruktiven Beitrag für die Unterrichts- und Schulentwicklung individueller Klassen und Schulen leisten.

Damit dies funktionieren kann, müssen solche Low-Stakes Assessment Systeme einen gemeinsamen Qualitätsstandard als Grundlage haben – nicht nur hinsichtlich der Messung von Schülerkompetenzen sondern auch hinsichtlich des methodischen Vorgehens bei der Auswertung und Rückmeldung von Testergebnissen. Ein verantwortungsvoller Umgang mit den Testergebnissen setzt voraus, dass die Analyse der Testdaten methodischen Mindestanforderungen genügt und transparent dargestellt wird. Die Heterogenität des Vorgehens bei der Datenanalyse in Vergleichsarbeiten zwischen den Bundesländern (vgl. Kapitel 4) spiegelt diesbezüglich einen Missstand wider.

Die vorliegende Arbeit zeigte auf, dass es nicht eine Standardprozedur bzw. das richtige Verfahren gibt, um fairere Vergleiche zu erstellen. Jedoch lassen sich inhaltliche und statistische Kriterien zur Beurteilung und Entwicklung fairerer Vergleiche ableiten, die als methodische Standards im Kontext von Adjustierungsverfahren in Vergleichsarbeiten dienen können. Die in dieser Arbeit betrachteten Kriterien der Fairness und Praktikabilität von Adjustierungsverfahren (vgl. Kapitel 4) können dabei als ein Element hinsichtlich der Ausdifferenzierung eines solchen gemeinsamen Standards fun-

gieren. Erste Schritte in diese Richtung liegen bspw. in Form der Untersuchung von Kuhl et al. (2011) sowie der vorliegenden Arbeit vor. Die Etablierung eines solchen gemeinsamen Standards wiederum kann die Umsetzung der KMK-Gesamtstrategie zum Bildungsmonitoring des deutschen Bildungssystems – insbesondere mit Blick auf die landesweiten Vergleichsarbeiten – zukünftig weiter unterstützen.

Das Potenzial fairerer Vergleiche besteht insbesondere darin, als externe Informationsbasis für die Lehrkräfte einer Klasse zu fungieren, die für die Unterrichtsentwicklung nutzbar gemacht werden kann. Dabei sollten die in Vergleichsarbeiten berechneten adjustierten Maße nicht als kausale Effekte des Unterrichts interpretiert werden. Dies trifft auf standardisierte Schulleistungstests, die im Kontext von Educational-Accountability-Systemen eingesetzt werden, im Allgemeinen zu. Gleichwohl können die faireren Vergleichswerte als deskriptive Maße Ausgangspunkt kollegialer Diskussionen sein sowie Impulse für Unterrichtsentwicklungsmaßnahmen geben. Zu diesem Zweck müssen den Rezipienten die Möglichkeiten und die Grenzen der Interpretation transparent sein.

> *„Causal inference may well be the holy grail of quantitative research in the social sciences, but it should not be proclaimed lightly. When the causal language of teacher 'effects' or 'effectiveness' is casually applied to the estimates from a value-added model simply because it conditions on a prior year test score, it trivializes the entire enterprise. And instead of promoting discussion among parents, teachers and school administrators about what students are and are not learning in their classrooms, it seems much more likely to shut them down."*
>
> – Briggs & Domingue (2011, S. 21) –

Literatur

Abayomi, K., Gelman, A. & Levy, M. (2008). Diagnostics for multivariate imputations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *57* (3), 273–291. doi: 10.1111/j.1467-9876.2007.00613.x

Aitkin, M. & Longford, N. (1986). Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *149* (1), 1–43. doi: 10.2307/2981882

Allison, P. D. (2001). *Missing data*. Thousand Oaks, CA: Sage Publications.

American Association for the Advancement of Science (Hrsg.). (1993). *Benchmarks for science literacy. Project 2061*. New York, NY: Oxford University Press.

Arbeit, C., Canter, M., Dabrowski, A., Dailey, D. & Oberman, I. (2007). *Balancing act: Best practices in the middle grades. A report from the California Best Practices Study*. San Francisco, CA: Springboard Schools.

Ballou, D., Sanders, W. & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics*, *29* (1), 37–65. doi: 10.3102/10769986029001037

Baraldi, A. N. & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology*, *48* (1), 5–37. doi: 10.1016/j.jsp.2009.10.001

Barnard, G. A. (1982). Causation. In S. Kotz, N. Johnson & C. Read (Hrsg.), *Encyclopedia of Statistical Sciences* (Bd. 1, S. 387–389). New York, NY: John Wiley & Sons.

Bauer, H. (1990). *Maß- und Integrationstheorie*. Berlin: de Gruyter.

Bauer, H. (2001). *Wahrscheinlichkeitstheorie* (5. Aufl.). Berlin: de Gruyter.

Baumert, J. & Artelt, C. (2003). Konzeption und technische Grundlagen der Studie. In Deutsches PISA-Konsortium (Hrsg.), *PISA 2000 – Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland* (S. 11–50). Opladen: Leske+Budrich.

Baumert, J., Becker, M., Neumann, M. & Nikolova, R. (2009). Frühübergang in ein grundständiges Gymnasium – Übergang in ein privilegiertes Entwicklungsmilieu? *Zeitschrift für Erziehungswissenschaft*, *12* (2), 189–215. doi: 10.1007/s11618-009-0072-4

Baumert, J., Bos, W. & Lehmann, R. (Hrsg.). (2000a). *Dritte Internationale Mathematik- und Naturwissenschaftsstudie: Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn. Bd. 1: Mathematisch-naturwissenschaftliche Grundbildung am Ende der Pflichtschulzeit*. Opladen: Leske+Budrich.

Baumert, J., Bos, W. & Lehmann, R. (Hrsg.). (2000b). *Dritte Internationale Mathematik- und Naturwissenschaftsstudie: Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn. Bd. 2: Mathematische und physikalische Kompetenzen am Ende der gymnasialen Oberstufe*. Opladen: Leske+Budrich.

Baumert, J., Lehmann, R., Lehrke, M., Schmitz, B., Clausen, M., Hosenfeld, I., … Neubrand, J. (1997). *TIMSS – Mathematisch-naturwissenschaftlicher Unterricht im internationalen Vergleich*. Opladen: Leske+Budrich.

Baumert, J. & Schümer, G. (2001). Familiäre Lebensverhältnisse, Bildungsbeteiligung und Kompetenzerwerb. In Deutsches PISA-Konsortium (Hrsg.), *PISA 2000 – Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (S. 323–410). Opladen: Leske+Budrich.

Baumert, J., Stanat, P. & Watermann, R. (2006). Schulstruktur und die Entstehung differenzieller Lern- und Entwicklungsmilieus. In J. Baumert, P. Stanat & R. Watermann (Hrsg.), *Herkunftsbedingte Disparitäten im Bildungswesen: Differenzielle Bildungsprozesse und Probleme der Verteilungsgerechtigkeit* (S. 95–188). Wiesbaden: VS Verlag für Sozialwissenschaften.

Benton, T., Hutchinson, D., Schagen, I. & Scott, E. (2003). *Study of the performance of maintained secondary schools in england*. Slough: National Foundation for Educational Research. Zugriff auf `http://www.leeds.ac.uk/educol/documents/00003494.htm`

Blalock, H. M. (1984). Contextual-effects models: Theoretical and methodological issues. *Annual Review of Sociology*, *10* (1), 353–372. doi: 10.1146/annurev.so.10.080184.002033

Blossfeld, H.-P., Bos, W., Daniel, H.-D., Hannover, B., Lenzen, D., Prenzel, M. & Wöß-

mann, L. (2010). *Bildungsautonomie: Zwischen Regulierung und Eigenverantwortung – Jahresgutachten 2010.* Wiesbaden: VS Verlag für Sozialwissenschaften.

Blum, W., Drüke-Noe, C., Hartung, R. & Köller, O. (Hrsg.). (2006). *Bildungsstandards Mathematik: Konkret – Sekundarstufe I: Aufgabenbeispiele, Unterrichtsanregungen, Fortbildungsideen*. Berlin: Cornelsen-Scriptor-Verlag.

Bollen, K. A. (1989). *Structural equation modeling with latent variables*. New York, NY: John Wiley & Sons.

Bonsen, M., Bos, W., Gröhlich, C., Harney, B., Imhäuser, K., Makles, A., ... Wendt, H. (2010). Zur Konstruktion von Sozialindizes – Ein Beitrag zur Analyse sozialräumlicher Benachteiligung von Schulen als Voraussetzung für qualitative Schulentwicklung. In Bundesministerium für Bildung und Forschung (Hrsg.), *Bildungsforschung* (Bd. 31). Berlin: BMBF.

Bos, W., Bonsen, M. & Gröhlich, C. (2010). *KESS 7 – Kompetenzen und Einstellungen von Schülerinnen und Schülern an Hamburger Schulen zu Beginn der Jahrgangsstufe 7*. Münster: Waxman.

Bos, W., Gröhlich, C. & Bonsen, M. (2010). Der Belastungsindex für die Schulen der Sekundarstufe I in Hamburg. In W. Bos, M. Bonsen & C. Gröhlich (Hrsg.), *KESS 7 – Kompetenzen und Einstellungen von Schülerinnen und Schülern an Hamburger Schulen zu Beginn der Jahrgangsstufe 7* (S. 123–131). Münster: Waxmann.

Bos, W., Lankes, E. M., Prenzel, M., Schwippert, K., Valtin, R. & Walther, G. (Hrsg.). (2003). *Erste Ergebnisse aus IGLU. Schülerleistungen am Ende der vierten Jahrgangsstufe im internationalen Vergleich*. Münster: Waxmann.

Bos, W. & Pietsch, M. (Hrsg.). (2006). *KESS 4 - Kompetenzen und Einstellungen von Schülerinnen und Schülern am Ende der Jahrgangsstufe 4 in Hamburger Grundschulen*. Münster: Waxmann.

Bos, W., Pietsch, M., Gröhlich, C. & Janke, N. (2006). Ein Belastungsindex für Schulen als Grundlage der Ressourcenzuweisung am Beispiel von KESS 4. Versuch einer Klassifizierung von Schultypen. In G. Holtappels, W. Bos, H. Pfeiffer, H.-G. Rolff & R. Schulz-Zander (Hrsg.), *Jahrbuch der Schulentwicklung.* Weinheim: Juventa.

Bos, W. & Schwippert, K. (2002). TIMSS, PISA, IGLU & Co. Vom Sinn und Unsinn internationaler Schulleistungsuntersuchungen. *Bildung und Erziehung*, *55* (1), 5–23.

Bourdieu, P. (1983). Ökonomisches Kapital, kulturelles Kapital, soziales Kapital. In R. Kreckel (Hrsg.), *Soziale Ungleichheiten* (S. 183–198). Göttingen: Schwartz.

Box, G. E. P. & Draper, N. R. (1986). *Empirical model-building and response surface*. New York, NY: John Wiley & Sons.

Bracey, G. W. (2005). *No child left behind: Where does the money go?* (Policy Report No. EPSL-0506-114-EPRU). Tempe, AZ: Arizona State University. Education Policy Studies Laboratory. Zugriff auf `http://nepc.colorado.edu/files/`

Braun, H., Chudowsky, N. & Koenig, J. (Hrsg.). (2010). *Getting value out of value-added: Report of a workshop*. Washington, DC: National Academies Press.

Braun, H. & Wainer, H. (2007). Value-added modeling. In C. R. Rao & S. Sinharay (Hrsg.), *Handbook of statistics: Vol. 26. Psychometrics* (S. 867–892). Amsterdam: Elsevier.

Breiman, L. (2001). Random forests. *Machine Learning*, *45* (1), 5–32. doi: 10.1023/A:1010933404324

Briggs, D. C. (2008). *The goals and uses of value-added models*. Paper prepared for the workshop of the Committee on Value-Added Methodology for Instructional Improvement, Program Evaluation, and Educational Accountability, National Research Council, Washington, DC.

Briggs, D. C. & Domingue, B. (2011). *Due diligence and the evaluation of teachers: A review of the value-added analysis underlying the effectiveness rankings of Los Angeles unified school district teachers by the Los Angeles Times*. Boulder, CO: National Education Policy Center.

Briggs, D. C., Weeks, J. P. & Wiley, E. (2009). The sensitivity of value-added modeling to the creation of a vertical score scale. *Education Finance and Policy*, *4* (4), 384–414. doi: 10.1162/edfp.2009.4.4.384

Briggs, D. C. & Wiley, E. W. (2008). Causes and effects. In K. E. Ryan & L. A. Shepard (Hrsg.), *The future of test-based educational accountability.* New York, NY: Routledge.

Bryk, A. S. & Weisberg, H. I. (1976). Value-added analysis: A dynamic approach to the estimation of treatment effects. *Journal of Educational and Behavioral Statistics*, *1* (2), 127–155. doi: 10.3102/10769986001002127

Bundesministerium für Bildung und Forschung (Hrsg.). (2008). *Wissen für Handeln – Forschungsstrategien für eine evidenzbasierte Bildungspolitik*. Berlin: BMBF.

Ceci, S. J. & Liker, J. K. (1986). A day at the races: A study of IQ, expertise, and

cognitive complexity. *Journal of Experimental Psychology: General*, *115* (3), 255–266. doi: 10.1037/0096-3445.115.3.255

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2. Aufl.). Hillsdale, NJ: Erlbaum.

Collins, L. M., Schafer, J. L. & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6* (4), 330–351. doi: 10.1037//1082-989X.6.4.330

Cronbach, L. J. (1976). *Research on classrooms and schools: Formulations of questions, designs and analysis*. Stanford, CA: Stanford Evaluation Consortium.

Department for Education. (2010). *The importance of teaching: The schools white paper 2010*. Zugriff auf `https://www.education.gov.uk/publications/`

Deutscher Bildungsrat. (1974). *Aspekte für die Planung der Bildungsforschung. Empfehlungen der Bildungskommission*. Stuttgart: Klett.

Deutsches PISA-Konsortium (Hrsg.). (2001). *PISA 2000 – Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich*. Opladen: Leske+Budrich.

Deutsches PISA-Konsortium (Hrsg.). (2002). *PISA 2000 – Die Länder der Bundesrepublik Deutschland im Vergleich*. Opladen: Leske+Budrich.

Ditton, H. (2000). Qualitätskontrolle und -sicherung in Schule und Unterricht – Ein Überblick zum Stand der empirischen Forschung [Zeitschrift für Pädagogik, 41. Beiheft]. In A. Helmke, W. Hornstein & E. Terhart (Hrsg.), *Qualitätssicherung im Bildungsbereich.* Weinheim: Beltz.

Ehmke, T. & Siegle, T. (2005). ISEI, ISCED, HOMEPOS, ESCS. Indikatoren der sozialen Herkunft bei der Quantifizierung von sozialen Disparitäten. *Zeitschrift für Erziehungswissenschaft*, *8* (4), 521–539. doi: 10.1007/s11618-005-0157-7

Elstrodt, J. (2009). *Maß- und Integrationstheorie* (6. Aufl.). Berlin: Springer.

Emmrich, R., Ernst, C.-M., Harych, P. & Wesselhöfft, K. (2012). *Vergleichsarbeiten in der Jahrgangsstufe 8 in Berlin als Beitrag zur Schul- und Unterrichtsentwicklung.* Institut für Schulqualität der Länder Berlin und Brandenburg. Zugriff auf `http://www.isq-bb.de/`

Enders, C. K. & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, *12* (2), 121–138. doi: 10.1037/1082-989X.12.2.121

Evans, H. (2008). *Value-added in english schools*. Paper presented at the National

Conference on Value-added Modeling, Madison, WI.

Evans, M. D., Kelley, J., Sikora, J. & Treiman, D. J. (2010). Family scholarly culture and educational success: Books and schooling in 27 nations. *Research in Social Stratification and Mobility*, *28* (2), 171–197. doi: 10.1016/j.rssm.2010.01.002

Fahrmeier, L., Kneib, T. & Lang, S. (2007). *Regression: Modelle, Methoden und Anwendungen*. Berlin: Springer.

Felch, J., Song, J. & Smith, D. (2010, 14. August). *Who's teaching L.A.'s kids?* Los Angeles Times. Zugriff auf `http://www.latimes.com/news/local/la-me-teachers-value-20100815,0,2695044.story`

Fiege, C. (2007). *Faire Vergleiche in Schulleistungsuntersuchungen und ihre kausaltheoretische Grundlage*. Unveröffentlichte Diplomarbeit, Friedrich-Schiller-Universität Jena.

Fiege, C. (in Druck). Faire Vergleiche bei Vergleichsarbeiten: Möglichkeiten und Grenzen. In B. Groot-Wilken, K. Isaac & J.-P. Schräpler (Hrsg.), *Sozialindex: Modelle und Anwendungsgebiete.* Münster: Waxmann.

Fiege, C., Reuther, F. & Nachtigall, C. (2011). Faire Vergleiche? Berücksichtigung von Kontextbedingungen des Lernens beim Vergleich von Testergebnissen aus deutschen Vergleichsarbeiten. *Zeitschrift für Bildungsforschung*, *1* (2), 133–149. doi: 10.1007/s35834-011-0009-x

Fisher, R. A. (1946). *Statistical methods for research workers* (10. Aufl.). Edinburgh: Oliver and Boyd.

Geneletti, S. & Dawid, A. P. (2011). Defining and identifying the effect of treatment on the treated. In P. M. Illari, F. Russo & J. Williamson (Hrsg.), *Causality in the sciences* (S. 728–749). New York, NY: Oxford University Press.

Georgii, H.-O. (2007). *Stochastik – Einführung in die Wahrscheinlichkeitstheorie und Statistik* (3. Aufl.). Berlin: de Gruyter.

Goldhaber, D. D., Goldschmidt, P. & Tseng, F. (2013). Teacher value-added at the high-school level: Different models, different answers? *Educational Evaluation and Policy Analysis*, first published online on January 16, 2013. doi: 10.3102/0162373712466938

Goldstein, H. (1997). Methods in school effectiveness research. *School effectiveness and school improvement*, *8* (4), 369–395. doi: 10.1080/0924345970080401

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60*, 549–576. doi: 10.1146/annurev.psych.58.110405

.085530

Graham, J. W., Olchowski, A. E. & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science*, *8*, 206–213. doi: 10.1007/s11121-007-0070-9

Gräsel, C. (2011). Was ist Empirische Bildungsforschung? In H. Reinders, H. Ditton, C. Gräsel & B. Gniewosz (Hrsg.), *Empirische Bildungsforschung: Strukturen und Methoden* (S. 13–37). Wiesbaden: VS Verlag für Sozialwissenschaften.

Gräsel, C. & Göbel, K. (2011). Unterrichtsqualität. In H. Reinders, H. Ditton, C. Gräsel & B. Gniewosz (Hrsg.), *Empirische Bildungsforschung: Gegenstandsbereiche* (S. 87–97). Wiesbaden: VS Verlag für Sozialwissenschaften.

Hamilton, L. S., Stecher, B. M., Marsh, J. A., McCombs, J. S., Robyn, A., Russell, J. L., … Barney, H. (2007). *Standards-based accountability under No Child Left Behind: Experiences of teachers and administrators in three states*. Santa Monica, CA: RAND Corporation.

Harker, R. & Tymms, P. (2004). The effects of student composition on school outcomes. *School Effectiveness and School Improvement*, *15* (2), 177–199. doi: 10.1076/sesi.15.2.177.30432

Hartig, J. & Frey, A. (2012). Konstruktvalidierung und Skalenbeschreibung in der Kompetenzdiagnostik durch die Vorhersage von Aufgabenschwierigkeiten. *Psychologische Rundschau*, *63* (1), 43–49. doi: 10.1026/0033-3042/a000109

Hartig, J. & Klieme, E. (2006). Kompetenz und Kompetenzdiagnostik. In K. Schweizer (Hrsg.), *Leistung und Leistungsdiagnostik* (S. 127–143). Heidelberg: Springer.

Hartig, J., Klieme, E. & Leutner, D. (2008). *Assessment of competencies in educational settings: State of the art and future prospects*. Göttingen: Hogrefe.

Hattie, J. A. (2002). Classroom composition and peer effects. *International Journal of Educational Research*, *37* (5), 449–481. doi: 10.1016/S0883-0355(03)00015-6

Hausknecht, H. & Eyrainer, J. (2010). *Ländergemeinsame Vergleichsarbeiten in Bayern – VERA-8: Handreichung für die Umsetzung an bayerischen Schulen.* Bayerisches Staatsministerium für Unterricht und Kultus. Zugriff auf `http://vergleichsarbeiten.isb-qa.de`

Heckman, J. J. & Robb, R. (1985). Alternative methods for evaluating the impact of interventions: An overview. *Journal of Econometrics*, *30* (1), 239–267. doi: 10.1016/0304-4076(85)90139-3

Hedges, L. V. & Hedberg, E. C. (2007). Intraclass correlation values for planning group-

randomized trials in education. *Educational Evaluation and Policy Analysis*, *29* (1), 60–87. doi: 10.3102/0162373707299706

Heller, K. A. & Hany, E. A. (2001). Standardisierte Schulleistungsmessungen. In F. E. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (S. 87–101). Weinheim: Beltz.

Heller, K. A. & Perleth, C. (2000). *Kognitiver Fähigkeitstest für 4. bis 12. Klassen, Revision (KFT 4-12+ R)*. Göttingen: Beltz-Test GmbH.

Helmke, A. & Hosenfeld, I. (2004). Vergleichsarbeiten – Kompetenzmodelle – Standards. In M. Wosnitza, A. Frey & R.-S. Jäger (Hrsg.), *Lernprozesse, Lernumgebungen und Lerndiagnostik. Wissenschaftliche Beiträge zum Lernen im 21. Jahrhundert* (S. 56–75). Landau: Verlag Empirische Pädagogik.

Helmke, A. & Hosenfeld, I. (2005). Standardbasierte Unterrichtsevaluation. In B. Brägger, B. Bucher & N. Landwehr (Hrsg.), *Schlüsselfragen zur externen Schulevaluation* (S. 127–151). Bern: h.e.p.-Verlag.

Helmke, A., Hosenfeld, I. & Schrader, F.-W. (2004). Vergleichsarbeiten als Instrument zur Verbesserung der Diagnosekompetenz von Lehrkräften. In R. Arnold & C. Griese (Hrsg.), *Schulleitung und Schulentwicklung* (S. 119–144). Hohengehren: Schneider-Verlag.

Hernán, M. A. & Robins, J. M. (2006). Instruments for causal inference: An epidemiologist's dream? *Epidemiology*, *17* (4), 360–372. doi: 10.1097/01.ede.0000222409.00878.37

Hill, P. W. & Rowe, K. J. (1996). Multilevel modelling in school effectiveness research. *School Effectiveness and School Improvement*, *7* (1), 1–34. doi: 10.1080/0924345960070101

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81* (396), 945–960. doi: 10.1080/01621459.1986.10478354

Hotz, V. J., Imbens, G. W. & Klerman, J. A. (2006). Evaluating the differential effects of alternative welfare-to-work training components: A re-analysis of the California GAIN program. *Journal of Labor Economics*, *24* (3), 521–566. doi: 10.1086/505050

Hovestadt, G. & Kessler, N. (2005). 16 Bundesländer - eine Übersicht zu Bildungsstandards und Evaluationen. In G. Becker et al. (Hrsg.), *Friedrich Jahresheft: Bd. 23. Standards – Unterrichten zwischen Kompetenzen, zentralen Prüfungen und Vergleichsarbeiten* (S. 8–10). Seelze: Friedrich Verlag.

Institut für Bildungsmonitoring und Qualitätsentwicklung (Hrsg.). (2012). *KERMIT: Hinweise und Anregungen zur Nutzung von KERMIT für die Unterrichts- und Schulentwicklung*. Hamburg: Behörde für Schule und Berufsbildung. Zugriff auf https://www.lernstand.hamburg.de/index.php?option=com_content&view=article&id=123&Itemid=133

Institut zur Qualitätsentwicklung im Bildungswesen. (2008). *Kompetenzstufenmodell zu den Bildungsstandards im Fach Mathematik für den Mittleren Schulabschluss.* Zugriff auf http://www.iqb.hu-berlin.de/bista/ksm

Isaac, K. & Hosenfeld, I. (2008). Faire Ergebnisrückmeldungen bei Vergleichsarbeiten. In J. Ramseger & M. Wagener (Hrsg.), *Chancenungleichheit in der Grundschule: Ursachen und Wege aus der Krise* (S. 144–145). Wiesbaden: VS Verlag für Sozialwissenschaften.

Iturre, R. A. C. (2005). The relationship between school composition, school process and mathematics achievement in secondary education in Argentina. *International Review of Education*, *51* (2-3), 173–200. doi: 10.1007/s11159-005-1843-7

Iversen, G. R. (1991). *Contextual analysis*. Newbury Park, CA: Sage.

Klenke, A. (2008). *Wahrscheinlichkeitstheorie* (2. Aufl.). Berlin: Springer.

Klieme, E. (2002). *Was ist PISA-E?* Frankfurt am Main: Deutsches Institut für Internationale Pädagogische Forschung (DIPF). Zugriff auf http://www.bildungsserver.de/db/mlesen.html?Id=15534

Klieme, E. et al. (Hrsg.). (2010). *PISA 2009. Bilanz nach einem Jahrzehnt*. Münster: Waxman.

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., ... Vollmer, H. J. (2007). *Zur Entwicklung nationaler Bildungsstandards: Eine Expertise*. Bonn: BMBF.

Klieme, E., Döbert, H., van Ackeren, I., Bos, W., Klemm, K., Lehmann, R. H., ... Weiß, M. (2007). Vertiefender Vergleich der Schulsysteme ausgewählter PISA-Teilnehmerstaaten. In Bundesministerium für Bildung und Forschung (Hrsg.), *Bildungsforschung* (Bd. 2). Berlin: BMBF.

Klieme, E. & Hartig, J. (2008). Kompetenzkonzepte in den Sozialwissenschaften und im erziehungswissenschaftlichen Diskurs. In M. Prenzel, I. Gogolin & H.-H. Krüger (Hrsg.), *Sonderheft 8 der Zeitschrift für Erziehungswissenschaft: Kompetenzdiagnostik* (S. 11–29). Wiesbaden: VS Verlag für Sozialwissenschaften.

Klieme, E. & Leutner, D. (2006). Kompetenzmodelle zur Erfassung individueller Lern-

ergebnisse und zur Bilanzierung von Bildungsprozessen. Beschreibung eines neu eingerichteten Schwerpunktprogramms der DFG. *Zeitschrift für Pädagogik*, *52* (6), 876–903.

KMK (Hrsg.). (1997). *Grundsätzliche Überlegungen zu Leistungsvergleichen innerhalb der Bundesrepublik Deutschland. Konstanzer Beschluss*. Zugriff auf http://www.kmk.org/dokumentation/veroeffentlichungen-beschluesse/bildung-schule/qualitaetssicherung-in-schulen.html

KMK (Hrsg.). (2002). *Bildungsstandards zur Sicherung von Qualität und Innovation im föderalen Wettbewerb der Länder*. Zugriff auf http://www.kmk.org/dokumentation/veroeffentlichungen-beschluesse/bildung-schule/qualitaetssicherung-in-schulen.html

KMK (Hrsg.). (2004a). *Bildungsstandards im Fach Deutsch für den Mittleren Schulabschluss – Beschluss vom 4.12.2003*. München: Luchterhand.

KMK (Hrsg.). (2004b). *Bildungsstandards im Fach Mathematik für den Mittleren Schulabschluss – Beschluss vom 4.12.2003*. München: Luchterhand.

KMK (Hrsg.). (2005). *Bildungsstandards der Kultusministerkonferenz – Erläuterungen zur Konzeption und Entwicklung*. Neuwied: Luchterhand.

KMK (Hrsg.). (2006). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring*. München: LinkLuchterhand.

KMK (Hrsg.). (2012). *Vereinbarung zur Weiterentwicklung von VERA*. Zugriff auf http://www.kmk.org/dokumentation/veroeffentlichungen-beschluesse/bildung-schule/qualitaetssicherung-in-schulen.html

Köller, O. (2010). Bildungsstandards. In R. Tippelt & B. Schmidt (Hrsg.), *Handbuch Bildungsforschung* (3. Aufl., S. 529–548). Wiesbaden: VS Verlag für Sozialwissenschaften.

Köller, O. (2011). Standardsetzung im Bildungssystem . In H. Reinders, H. Ditton, C. Gräsel & B. Gniewosz (Hrsg.), *Empirische Bildungsforschung: Strukturen und Methoden* (S. 179–192). Wiesbaden: VS Verlag für Sozialwissenschaften.

Kolmogoroff, A. N. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Berlin: Springer.

Kreft, I. G., de Leeuw, J. & Aiken, L. S. (1995). The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research*, *30*

(1), 1–21. doi: 10.1207/s15327906mbr3001_1

Kuhl, P., Lenkeit, J., Pant, H. A. & Wendt, W. (2011). Die Kontextuierung von Leistungswerten bei Vergleichs- und Prüfungsarbeiten. Verschiedene Wege, die Zusammensetzung der Schülerschaft in den Rückmeldungen an Schulen und die Schulinspektion zu berücksichtigen. In S. Müller, M. Pietsch & W. Bos (Hrsg.), *Schulinspektion in Deutschland: Eine Zwischenbilanz aus empirischer Sicht* (S. 237–259). Münster: Waxmann.

Kuper, H. & Schneewind, J. (2006). *Rückmeldung und Rezeption von Forschungsergebnissen – Zur Verwendung wissenschaftlichen Wissens im Bildungssystem*. Münster: Waxmann.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of trainig programs with experimental data. *The American Economic Review*, *76* (4), 604–620. doi: 10.2307/1806062

Lechner, M. (2000). *A note on the common support problem in applied evaluation studies* (Discussion paper 2001-01). Department of Economics, University of St. Gallen, Switzerland. Zugriff auf `http://papers.ssrn.com/sol3/papers.cfm?abstract_id=259239`

Leckie, G. & Goldstein, H. (2009). The limitations of using school league tables to inform school choice. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *172* (4), 835–851. doi: 10.1111/j.1467-985X.2009.00597.x

Lehmann, R. & Lenkeit, J. (2008). *ELEMENT. Erhebung zum Lese und Mathematikverständnis. Entwicklung in den Jahrgangsstufen 4 bis 6 in Berlin: Abschlussbericht über die Untersuchungen 2003, 2004 und 2005 an Berliner Grundschulen und grundständigen Gymnasien*. Berlin: Humboldt Universität zu Berlin.

Leucht, M., Harsch, C., Pant, H. A. & Köller, O. (2012). Steuerung zukünftiger Aufgabenentwicklung durch Vorhersage der Schwierigkeiten eines Tests für die erste Fremdsprache Englisch durch Dutch Grid Merkmale. *Diagnostica*, *58* (1), 31–44. doi: 10.1026/0012-1924/a000063

Liaw, A. & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2* (3), 18–22.

Light, R. J. & Pillemer, D. (1984). *Summing up: The science of reviewing research*. Cambridge: Harvard University Press.

Lind, G. (2009). Amerika als Vorbild? Erwünschte und unerwünschte Folgen aus Evaluationen. In T. Bohl & H. Kiper (Hrsg.), *Lernen aus Evaluationsergebnissen*

– *Verbesserungen planen und implementieren* (S. 63–81). Bad Heilbrunn: Julius Klinkhardt.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher*, *29* (2), 4–16. doi: 10.3102/0013189X029002004

Linn, R. L., Baker, E. L. & Betebenner, D. W. (2002). Accountability systems: Implications of requirements of the No Child Left Behind Act of 2001. *Educational Researcher*, *31* (6), 3–16. doi: 10.3102/0013189X031006003

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, *83* (404), 1198–1202. doi: 10.1080/01621459.1988.10478722

Little, R. J. A. & Rubin, D. B. (2002). *Statistical Analysis with Missing Data* (2. Aufl.). Hoboken, NJ: Wiley.

Lorenz, J. H. (2005). Zentrale Lernstandsmessung in der Primarstufe: Vergleichsarbeiten Klasse 4 (VERA) in sieben Bundesländern. *Zentralblatt für Didaktik der Mathematik*, *37* (4), 317–323. doi: 10.1007/BF02655818

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T. & Muthén, B. O. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods*, *13* (3), 203–229. doi: 10.1037/a0012869

Lüdtke, O., Robitzsch, A., Trautwein, U. & Köller, O. (2007). Umgang mit fehlenden Werten in der psychologischen Forschung. Probleme und Lösungen. *Psychologische Rundschau*, *58* (2), 103–117. doi: 10.1026/0033-3042.58.2.103

Maaz, K., Baumert, J., Gresch, C. & McElvany, N. (2010). Der Übergang von der Grundschule in die weiterführende Schule: Leistungsgerechtigkeit und regionale, soziale und ethnisch-kulturelle Disparitäten. In Bundesministerium für Bildung und Forschung (Hrsg.), *Bildungsforschung* (Bd. 34). Berlin: BMBF.

Maaz, K., Trautwein, U. & Dumont, H. (2011). *Definition und Verteilung von Schulen mit benachteiligter Schülerschaft*. Expertise im Auftrag der Bertelsmann Stiftung.

Maier, U. (2008). Vergleichsarbeiten im Vergleich – Akzeptanz und wahrgenommener Nutzen standardbasierter Leistungsmessungen in Baden-Württemberg und Thüringen. *Zeitschrift für Erziehungswissenschaft*, *11*, 453–474. doi: 10.1007/s11618-008-0036-0

Maier, U. (2009). *Wie gehen Lehrerinnen und Lehrer mit Vergleichsarbeiten um? Ei-*

ne Studie zu testbasierten Schulreformen in Baden-Württemberg und Thüringen. Baltmannsweiler: Schneider Hohengehren.

Marsh, H. W. (1984). Self-concept: The application of a frame of reference model to explain paradoxical results. *Australian Journal of Education*, *28* (2), 165–181.

Marsh, H. W. (1987). The big-fish-little-pond effect on academic self-concept. *Journal of Educational Psychology*, *79* (3), 280–295. doi: 10.1037/0022-0663.79.3.280

Marsh, H. W., Lüdtke, O., Robitzsch, A., Trautwein, U., Asparouhov, T., Muthén, B. & Nagengast, B. (2009). Doubly-latent models of school contextual effects: Integrating multilevel and structural equation approaches to control measurement and sampling error. *Multivariate Behavioral Research*, *44* (6), 764–802. doi: 10.1080/00273170903333665

Mayer, A., Thoemmes, F., Rose, N. & Steyer, R. (2011). *Theory and analysis of total, direct and indirect causal effects.* Manuskript in Vorbereitung, Friedrich-Schiller-Universität Jena.

McArdle, J. J. & Epstein, D. (1987). Latent growth curves within developmental structural equation models. *Child Development*, *58* (1), 110–133. doi: 10.2307/1130295

McArdle, J. J. & Hamagami, F. (2001). Latent difference score structural models for linear dynamic analyses with incomplete longitudinal data. In L. M. Collins & A. G. Sayer (Hrsg.), *New methods for the analysis of change* (S. 139–175). Washington, DC: American Psychological Association.

McCaffrey, D. F., Lockwood, J., Koretz, D., Louis, T. A. & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*, *29* (1), 67–101. doi: 10.3102/10769986029001067

McCaffrey, D. F., Lockwood, J. R., Koretz, D. & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability*. Santa Monica, CA: RAND Corporation.

McCaffrey, D. F., Stuart, E. A., Rubin, D. B. & Zanutto, E. L. (2006). *Design and implementation of case-control matching to estimate the effects of value-added assessment*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

McCulloch, C. E. & Searle, S. R. (2001). *Generalized, linear, and mixed models*. New York, NY: Wiley.

Meredith, M. & Tisak, J. (1990). Latent curve analysis. *Psychometrika*, *55* (1), 107–

122. doi: 10.1007/BF02294746

Meyer, R. H. (1997). Value-added indicators of school performance: A primer. *Economics of Education Review*, *16*, 283–301. doi: 10.1016/S0272-7757(96)00081-7

Mill, J. S. (1865). Of the four methods of experimental inquiry. In J. S. Mill (Hrsg.), *A system of logic, ratiocinative and inductive: Being a connected view of the principles of evidence, and the methods of scientific investigation* (6. Aufl., Bd. 1, S. 427–450). London: Longmans, Green, and Co.

Müller, A. (2010). *Rückmeldungen nach Vergleichsarbeiten im Kontext des schulischen Qualitätsmanagements. Drei explorative Studien zu Gestaltung und Rezeption im Anschluss an KOALA-S*. Berlin: Mensch und Buch Verlag.

Mooney, C. Z. (1997). *Monte carlo simulation*. Thousands Oaks, CA: Sage Publications.

Moosbrugger, H. & Kelava, A. (2007). *Testtheorie und Fragebogenkonstruktion*. Heidelberg: Springer.

Morgan, S. L. & Harding, D. J. (2006). Matching estimators of causal effects: Prospects and pitfalls in theory and practice. *Sociological Methods & Research*, *35* (1), 3–60. doi: 10.1177/0049124106289164

Morgan, S. L. & Winship, C. (2007). *Counterfactuals and causal inference. Methods and principles for social research*. New York, NY: Cambridge University Press.

Nachtigall, C. (2008). *Landesbericht Thüringer Kompetenztests 2008*. Zugriff auf `https://www.kompetenztest.de/downloads/kompetenztests/archiv`

Nachtigall, C. (2010). *Landesbericht Thüringer Kompetenztests 2010*. Zugriff auf `https://www.kompetenztest.de/downloads/kompetenztests/archiv`

Nachtigall, C., Hempel, G., Jantowski, A., Kröhne, U. & Müller, M. (2004). *Landesbericht – Thüringer Kompetenztest 2004*. Zugriff auf `https://www.kompetenztest.de/downloads/kompetenztests/archiv`

Nachtigall, C. & Kröhne, U. (2003). *Ergänzende Hinweise zur Interpretation – Zur Berechnung „virtueller Vergleichsklassen“*. Zugriff auf `https://www.kompetenztest.de/downloads/kompetenztests/archiv`

Nachtigall, C. & Kröhne, U. (2006). Methodische Anforderungen an schulische Leistungsmessung – auf dem Weg zu fairen Vergleichen. In H. Kuper & J. Schneewind (Hrsg.), *Rückmeldung und Rezeption von Forschungsergebnissen* (S. 59–74). Münster: Waxman.

Nachtigall, C., Kröhne, U., Enders, U. & Steyer, R. (2008). Causal effects and fair

comparison: Considering the influence of context variables on student competencies. In J. Hartig, E. Klieme & D. Leutner (Hrsg.), *Assessment of competencies in educational contexts* (S. 297–316). New York: Hogrefe & Huber.

Nachtigall, C., Müller, M. & Storbeck, I. (2010). *Beispielbericht: Ergebnisbericht Mathematik Klasse 6.* Zugriff auf `https://www.kompetenztest.de/downloads/kompetenztests/archiv`

Nachtigall, C., Storbeck, I. & Landmann, M. (2009). Belastung oder Chance. Zur Nutzung von Vergleichsarbeiten, Lernstandserhebungen, Kompetenztests, Orientierungsarbeiten und Co. In *Schulleitung und Schulentwicklung.* 1–18.

Nagengast, B. (2009). *Causal inference in multilevel models*. Unveröffentlichte Dissertation, Friedrich-Schiller-Universität Jena.

National Council of Teachers of Mathematics (Hrsg.). (2000). *Professional standards for school mathematics*. Reston, VA: National Council of Teachers of Mathematics.

National Research Council (Hrsg.). (1995). *National science education standards*. Washington, DC: National Academy Press.

Netzwerk empiriegestützte Schulentwicklung. (2006). *Zentrale standardisierte Lernstandserhebungen* (Positionspapier von der 5. EMSE-Tagung in Berlin). Zugriff auf `http://www.emse-netzwerk.de/pmwiki.php/Main/Material`

Neumann, M., Schnyder, I., Trautwein, U., Niggli, A., Lüdtke, O. & Cathomas, R. (2007). Schulformen als differenzielle Lernmilieus. *Zeitschrift für Erziehungswissenschaft*, *10* (3), 399–420. doi: 10.1007/s11618-007-0043-6

Neyman, J. (1923/1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9 (D. M. Dabrowska & T. P. Speed, Übers..). *Statistical Science*, *5*, 465–480. (Originalarbeit erschienen 1923) doi: 10.1214/ss/1177012031

No Child Left Behind (NCLB) Act of 2001, Pub. L. No. 107-110, § 115, Stat 1425. (2002). Zugriff auf `http://www.ed.gov/legislation/ESEA02/`

Oberman, I. (2005). *Challenged schools, remarkable results: Three lessons from California's highest achieving high schools*. San Francisco, CA: Springboard Schools.

OECD. (2001). *Knowledge and skills for life: First results from PISA 2000.* Paris: OECD Publishing.

OECD. (2008). *Measuring improvements in learning outcomes: Best practices to*

assess the value-added of schools. Paris: OECD Publishing.

Oelkers, J. & Reusser, K. (2008). *Qualität entwickeln – Standards sichern – mit Differenz umgehen*. Berlin: BMBF.

Øksendal, B. (2007). *Stochastic differential equations: An introduction with applications* (6. Aufl.). Berlin: Springer.

Opdenakker, M.-C. & Van Damme, J. (2007). Do school context, student composition and school leadership affect school practice and outcomes in secondary education? *British Educational Research Journal*, *33* (2), 179–206. doi: 10.1080/01411920701208233

Orth, G. (2007). Lernstandserhebungen und zentrale Prüfungen. Zwei Königskinder, die zueinander kommen können? *Pädagogik*, *3*, 16–20.

Pant, H. A., Tiffin-Richards, S. P. & Köller, O. (2010). Standard-Setting für Kompetenztests im Large-Scale-Assessment. *Zeitschrift für Pädagogik*, *56*, 175–188.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge: Cambridge University Press.

Peugh, J. L. & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research*, *74* (4), 525–556. doi: 10.3102/00346543074004525

Popper, K. R. (1934/2005). *Logik der Forschung* (10. verb. u. verm. Aufl.). Tübingen: Mohr Siebeck.

R Development Core Team. (2010). R: A language and environment for statistical computing [Software-Handbuch]. Wien, Österreich. Zugriff auf `http://www.R-project.org`

Raghunathan, T. & Bondarenko, I. (2007). *Diagnostics for multiple imputations*. Zugriff auf `http://ssrn.com/abstract=1031750`

Raudenbush, S. W. (2004). What are value-added models estimating and what does this imply for statistical practice? *Journal of Educational and Behavioral Statistics*, *29* (1), 121–129. doi: 10.3102/10769986029001121

Raudenbush, S. W. & Bryk, A. S. (1986). A hierarchical model for studying school effects. *Sociology of Education*, *59* (1), 1–17.

Raudenbush, S. W. & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2. Aufl.). Thousands Oaks, CA: Sage Publications.

Raudenbush, S. W. & Willms, J. D. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics*, *20* (4), 307–335. doi: 10.3102/

10769986020004307

Ray, A. (2006). *A review of multilevel value-added models in education*. Paper presented at the OECD Project on the Development of Value-Added Models in Education Systems, England.

Raykov, T. (1999). Are simple change scores obsolete? An approach to studying correlates and predictors of change. *Applied Psychological Measurement*, *23* (2), 120–126. doi: 10.1177/01466219922031248

Reinders, H., Ditton, H., Gräsel, C. & Gniewosz, B. (Hrsg.). (2011). *Empirische Bildungsforschung: Strukturen und Methoden*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Renkl, A. (1996). Vorwissen und Schulleistung. In J. Möller & O. Köller (Hrsg.), *Emotionen, Kognitionen und Schulleistung* (S. 175–190). Weinheim: Beltz.

Rheinberg, F. (2001). Bezugsnormen und schulische Leistungsbeurteilung. In F. E. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (S. 59–71). Weinheim: Beltz.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, *15* (3), 351–357. doi: 10.2307/2087176

Rolff, H.-G. (2002). Rückmeldung und Nutzung der Ergebisse von großflächigen Leistungsuntersuchungen. Grenzen und Chancen. In R. Schulz-Zander (Hrsg.), *Jahrbuch der Schulentwicklung* (S. 75–98). Weinheim: Juventa.

Rose, N. (2013). *Item nonresponses in educational and psychological measurement*. Unveröffentlichte Dissertation, Friedrich-Schiller-Universität Jena.

Rosenbaum, P. R. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70* (1), 41–55. doi: 10.1093/biomet/70.1.41

Rowan, B., Correnti, R. & Miller, R. J. (2002). What large-scale survey research tells us about teacher effects on student achievement: Insights from the Prospects study of elementary schools. *Teachers College Record*, *104* (8), 1525–1567. doi: 10.1111/1467-9620.00212

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66* (5), 688–701. doi: 10.1037/h0037350

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63* (3), 581–590. doi: 10.1093/biomet/63.3.581

Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics*, *2* (1), 1–26. doi: 10.3102/10769986002001001

Rubin, D. B. (1978). Bayesian-inference for causal effects: The role of randomization. *Annals of Statistics*, *6* (1), 34–58. doi: 10.1214/aos/1176344064

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: Wiley.

Rubin, D. B. (2008a). Comment: The design and analysis of gold standard randomized experiments. *Journal of the American Statistical Association*, *103* (484), 1350–1353. doi: 10.1198/016214508000001011

Rubin, D. B. (2008b). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, *2* (3), 808–840. doi: 10.1214/08-AOAS187

Rubin, D. B., Stuart, E. A. & Zanutto, E. L. (2004). A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral Statistics*, *29* (1), 103–116. doi: 10.3102/10769986029001103

Rubinstein, R. Y. (1981). *Simulation and the Monte Carlo method*. New York, NY: Wiley.

Rumberger, R. W. & Palardy, G. J. (2005). Does segregation still matter? The impact of student composition on academic achievement in high school. *Teachers College Record*, *107* (9), 1999–2045.

Ryan, K. E. & Shepard, L. A. (2008). *The future of test-based educational accountability*. New York, NY: Routledge.

Sanders, W. L. & Horn, S. P. (1994). The Tennessee value-added assessment system (TVAAS): Mixed-model methods in educational assessment. *Journal of Personnel Evaluation in Education*, *8* (3), 299–311. doi: 10.1007/BF00973726

Sanders, W. L., Saxton, A. & Horn, S. P. (1997). The Tennessee value-added assessment system: A quantitative outcomes-based approach to educational assessment. In J. Millman (Hrsg.), *Grading teachers, grading schools: Is student achievement a valid evaluational measure?* (S. 137–162). Thousand Oaks, CA: Corwin Press.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London, Great Britain: Chapman and Hall.

Schafer, J. L. (1999). NORM: Multiple imputation of incomplete multivariate data under a normal model [Computer software]. University Park, PA: Department of Statistics, Pennsylvania State University.

Schafer, J. L. & Graham, J. W. (2002). Missing data: Our view of the state of the art.

Psychological Methods, *7* (2), 147–177. doi: 10.1037//1082-989X.7.2.147

Scheerens, J. (1990). School effectiveness research and the development of process indicators of school functioning. *School Effectiveness and School Improvement*, *1* (1), 61–80. doi: 10.1080/0924345900010106

Scheerens, J. (2008). *Review and meta-analyses of school and teaching effectiveness*. Berlin: BMBF.

Schneider, W. & Bjorklund, D. F. (1992). Expertise, aptitude, and strategic remembering. *Child Development*, *63* (2), 461–473. doi: 10.1111/j.1467-8624.1992.tb01640.x

Schrader, F.-W. & Helmke, A. (2008). Determinanten der Schulleistung. In M. Schweer (Hrsg.), *Lehrer-Schüler-Interaktion: Inhaltsfelder, Forschungsperspektiven und methodische Zugänge* (2. Aufl., S. 285–302). Wiesbaden: VS Verlag für Sozialwissenschaften.

Sengewald, M.-A. (2011). *Die Verwendung des fachspezifischen Vorwissens für die Berechnung von sozialen Bezugsnormen: Vergleich zweier Parametrisierungsansätze des Allgemeinen Linearen Modells*. Unveröffentlichte Diplomarbeit, Friedrich-Schiller-Universität Jena.

Shadish, W. R., Clark, M. H. & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association*, *103* (484), 1334–1344. doi: 10.1198/016214508000000733

Shpitser, I. & Pearl, J. (2009). Effects of treatment on the treated: Identification and generalization. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence* (S. 514–521). Arlington, VA: AUAI Press.

Snijders, T. A. B. & Bosker, R. J. (1999). *Multilevel analysis. An introduction to basic and advanced multilevel modeling*. Thousands Oaks, CA: Sage Publications.

Spiewak, M. (2011, 2. Juni). *„So war der Test nicht gemeint“: Die Vergleichsarbeiten (Vera) an Grundschulen sind in die Kritik geraten. Ein Gespräch mit Hans Anand Pant, der sie entwickelt hat.* DIE ZEIT. Zugriff auf `http://www.zeit.de/2011/22/C-Interview-Pant`

Steiner, P. M., Cook, T. D., Shadish, W. R. & Clark, M. H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods*, *15* (3), 250–267. doi: 10.1037/a0018719

Steyer, R. (1989). Models of classical psychometric test theory as stochastic measu-

rement models: Representation, uniqueness, meaningfulness, identifiability and testability. *Methodika*, *3*, 25–60.

Steyer, R. (1992). *Theorie kausaler Regressionsmodelle*. Stuttgart: Fischer.

Steyer, R. (2001). Classical test theory. In C. Ragin & T. D. Cook (Hrsg.), *International encyclopedia of the social and behavioural sciences: Logic of inquiry and research design* (S. 481–520). Oxford: Pergamon.

Steyer, R. (2003). *Wahrscheinlichkeit und Regression*. Berlin: Springer.

Steyer, R., Eid, M. & Schwenkmezger, P. (1997). Modeling true intraindividual change: True change as a latent variable. *Methods of Psychological Research Online*, *2* (1), 21–33.

Steyer, R., Gabler, S., von Davier, A. A. & Nachtigall, C. (2000). Causal regression models II: Unconfoundedness and causal unbiasedness. *Methods of Psychological Research Online*, *5* (3), 55–87.

Steyer, R., Gabler, S., von Davier, A. A., Nachtigall, C. & Buhl, T. (2000). Causal regression models I: Individual and average causal effects. *Methods of Psychological Research Online*, *5* (2), 39–71.

Steyer, R., Nachtigall, C., Wüthrich-Martone, O. & Kraus, K. (2002). Causal regression models III: Covariates, conditional, and unconditional average causal effects. *Methods of Psychological Research Online*, *7* (1), 41–68.

Steyer, R., Nagel, W., Partchev, I. & Mayer, A. (in Druck). *Probability and regression*. Heidelberg: Springer.

Steyer, R., Partchev, I., Kröhne, U., Nagengast, B. & Fiege, C. (2011). *Probability and causality*. Manuskript in Vorbereitung.

Steyer, R., Partchev, I., Kröhne, U., Nagengast, B. & Fiege, C. (in Druck). *Probability and causality*. Heidelberg: Springer.

Stuart, E. A. (2004). *Matching methods for estimating causal effects using multiple control groups*. Unveröffentlichte Dissertation, Department of Statistics, Harvard University.

Tekwe, C. D., Carter, R. L., Ma, C.-X., Algina, J., Lucas, M. E., Roth, J., ... Resnick, M. B. (2004). An empirical comparison of statistical methods for value-added assessment of school performance. *Journal of Educational and Behavioral Statistics*, *29* (1), 11–36. doi: 10.3102/10769986029001011

Templ, M. & Alfons, A. (2009, November). *Visualisierung von fehlenden Werten*. Vortrag auf dem 18. wissenschaftlichen Kolloquium „Informationsvisualisierung

- Grafische Aufbereitung und Analyse von statistischen Daten, Wiesbaden.

Templ, M., Alfons, A. & Kowarik, A. (2011). VIM: Visualization and imputation of missing values [Software-Handbuch]. Zugriff auf `http://cran.r-project.org/package=VIM` (R package version 2.0.1)

Thompson, B. (1999). *Common methodology mistakes in educational research, revisited, along with a primer on both effect sizes and the bootstrap*. Invited address presented at the annual meeting of the American Educational Research Association, Montreal.

Timmermans, A. C., Doolaard, S. & de Wolf, I. (2011). Conceptual and empirical differences among various value-added models for accountability. *School Effectiveness and School Improvement*, *22* (4), 393–413. doi: 10.1080/09243453.2011.590704

Tippelt, R. & Schmidt, B. (Hrsg.). (2010). *Handbuch Bildungsforschung* (3. Aufl.). Wiesbaden: VS Verlag für Sozialwissenschaften.

van Ackeren, I. (2002). Von FIMS und FISS bis TIMSS und PISA. Schulleistungen in Deutschland im historischen und internationalen Vergleich. *Die Deutsche Schule*, *94* (2), 157–175.

van Ackeren, I. (2003a). *Evaluation, Rückmeldung und Schulentwicklung. Erfahrungen mit zentralen Tests, Prüfungen und Inspektionen in England, Frankreich und den Niederlanden*. Münster: Waxman.

van Ackeren, I. (2003b). Nutzung großflächiger Tests für die Schulentwicklung. Erfahrungen aus England, Frankreich und den Niederlanden. In Bundesministerium für Bildung und Forschung (Hrsg.), *Reihe Bildungsreform* (Bd. 3). Berlin: BMBF.

van Ackeren, I. & Bellenberg, G. (2004). Parallelarbeiten, Vergleichsarbeiten und Zentrale Abschlussprüfungen. In H. G. Rolff, H. G. Holtappels, K. Klemm, H. Pfeiffer & R. Schulz-Zander (Hrsg.), *Jahrbuch der Schulentwicklung* (S. 125–159). Weinheim: Juventa.

van Ackeren, I. & Klemm, K. (2009). Die entwicklungsorientierte Perspektive: Wie können Schule und Unterricht durch Evaluation entwickelt werden? In I. van Ackeren & K. Klemm (Hrsg.), *Entstehung, Struktur und Steuerung des deutschen Schulsystems: Eine Einführung* (S. 155–180). Wiesbaden: VS Verlag für Sozialwissenschaften.

van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, *16* (3), 219–

242. doi: 10.1177/0962280206074463

van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M. & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, *76* (12), 1049–1064. doi: 10.1080/10629360600810434

van Buuren, S. & Groothuis-Oudshoorn, K. (2011a). MICE: Multivariate imputation by chained equations [Software-Handbuch]. Zugriff auf `http://cran.r-project.org/web/packages/mice/` (R package version 2.3)

van Buuren, S. & Groothuis-Oudshoorn, K. (2011b). MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45* (3), 1–67.

van Buuren, S. & Oudshoorn, C. G. M. (1999). *Flexible multivariate imputation by MICE* (TNO report PG/VGZ/99.054). Leiden, Niederlande: TNO Prevention Center.

Van Ewijk, R. & Sleegers, P. (2010). The effect of peer socioeconomic status on student achievement: A meta-analysis. *Educational Research Review*, *5* (2), 134–150. doi: 10.1016/j.edurev.2010.02.001

Wacker, A. & Kramer, J. (2012). Vergleichsarbeiten in Baden-Württemberg – Zur Einschätzung von Lehrkräften vor und nach der Implementation. *Zeitschrift für Erziehungswissenschaft*, *15* (4), 683–706. doi: 10.1007/s11618-012-0326-4

Watermann, R. & Stanat, P. (2004). Schulrückmeldung in PISA 2000: Sozialnorm- und kriteriumsorientierte Rückmeldeverfahren. *Empirische Pädagogik*, *18* (1), 40–61.

Watermann, R., Stanat, P., Kunter, M., Klieme, E. & Baumert, J. (2003). Schulrückmeldungen im Rahmen von Schulleistungsuntersuchungen: Das Disseminationskonzept von PISA-2000. *Zeitschrift für Pädagogik*, *49* (1), 92–111.

Wegscheider, K. (2004). Methodische Anforderungen an Einrichtungsvergleiche (Profiling) im Gesundheitswesen. *Zeitschrift für ärztliche Fortbildung und Qualität im Gesundheitswesen*, *98* (8), 647–654.

Weinert, F. E. (2002). *Leistungsmessungen an Schulen*. Weinheim: Beltz.

Weinert, F. E. & Helmke, A. (1995). Interclassroom differences in instructional quality and interindividual differences in cognitive development. *Educational Psychologist*, *30* (1), 15–20. doi: 10.1207/s15326985ep3001_2

Wendt, H. & Bos, W. (2011). Indikatoren zur Kontextuierung von Inspektionsergebnissen – Bedeutung und Anforderungen. In S. Müller, M. Pietsch & W. Bos (Hrsg.),

Schulinspektion in Deutschland: Eine Zwischenbilanz aus empirischer Sicht (S. 237–259). Münster: Waxman.

West, S. G. (2001). New approaches to missing data in psychological research: Introduction to the special section. *Psychological Methods*, *6* (4), 315–316. doi: 10.1037/1082-989X.6.4.315

Willms, J. D. (2008). *Seven key issues for assessing 'value added' in education*. Paper prepared for the workshop of the Committee on Value-Added Methodology for Instructional Improvement, Program Evaluation, and Educational Accountability, National Research Council, Washington, DC.

Willms, J. D. (2010). School composition and contextual effects on student outcomes. *Teachers College Record*, *112* (4), 1008–1037.

Willms, J. D. & Raudenbush, S. W. (1989). A longitudinal hierarchical linear model for estimating school effects and their stability. *Journal of Educational Measurement*, *26* (3), 209–232. doi: 10.1111/j.1745-3984.1989.tb00329.x

Winship, C. & Morgan, S. L. (1999). The estimation of causal effects from observational data. *Annual Review of Sociology*, *25*, 659–707. doi: 10.1146/annurev.soc.25.1.659

Yuan, K.-H. & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology*, *30* (1), 165–200. doi: 10.1111/0081-1750.00078

A Abkürzungsverzeichnis

Akronyme

ACE	durchschnittlicher kausaler Effekt (engl.: average causal effect)
AYP	adequate yearly progress
BFLPE	Big-Fish-Little-Pond Effect
CAM	Contextualized Attainment Model
CCE	bedingter kausaler Effekt (engl.: conditional causal effect)
CBPS	California Best Practices Study
CM	Contextual Model
CVA	Contextual Value-Added Model
EMSE	Netzwerk Empiriegestützte Schulentwicklung
EVAS	Eigenverantwortliche Schule
E.U.LE.	Entwicklungsprogramm für Unterricht und Lernqualität
FIML	Full Information Maximum Likelihood
GCSE	General Certificate of Secondary Education
IEA	International Association for the Evaluation of Educational Achievement
IGLU	Internationale Grundschul-Lese-Untersuchung (Nationale Bezeichnung von PIRLS in Deutschland)
IGLU-E	Nationale Erweiterung von IGLU in der Bundesrepublik Deutschland
i.i.d.	unabhängig und identisch verteilt (engl.: independent and identically distributed)
IncMSE	Veränderung des mittleren Fehlerquadrats (engl.: increase in mean square error)
IQB	Institut zur Qualitätsentwicklung im Bildungswesen
ISB	Staatsinstitut für Schulqualität und Bildungsforschung (Bayern)
ISQ	Institut für Schulqualität der Länder Berlin und Brandenburg e.V.

KERMIT	Kompetenzen ermitteln
KFT	Kognitiver Fähigkeitstest (Heller & Perleth, 2000)
KMK	Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (Kurzform: Kultusministerkonferenz)
LAL 7	Ermittlung der Lernausgangslagen in Klassenstufe 7
LAUSD	Los Angeles Unified School District
LAVAM	Los Angeles Value-Added Model
MAR	Missing at random
MCAR	Missing completely at random
MI	Multiple Imputation
MICE	Multiple Imputation by Chained Equations
MNAR	Missing not at random
MSA	Mittlerer Schulabschluss
NCLB	No Child Left Behind
NEAP	National Assessment of Educational Progress
NPD	National Pupil Database
OECD	Organisation for Economic Cooperation and Development (Organisation für wirtschaftliche Zusammenarbeit und Entwicklung)
PFE	Prima-Facie-Effekt
PIRLS	Progress in International Reading Literacy Study
PISA	Programme for International Student Assessment
PISA-E	Nationale Erweiterung von PISA in der Bundesrepublik Deutschland
PLASC	Pupil Level Annual School Census
RR	Rechenregel
SES	Sozioökonomischer Status (engl.: socio-economic status)
TIMSS	Third International Mathematics and Science Study (ursprünglich); Trends in International Mathematics and Science Study (seit 2003)
TMBWK	Thüringer Ministerium für Bildung, Wissenschaft und Kultur
TOSCA	Transformation des Sekundarschulsystems und akademische Karrieren
TVAAS	Tennessee Value-Added Assessment System
UK	United Kingdom
VAM	Value-Added Model

Variablen

Y	Outcome-Variable
X	Treatment-Variable
Z	Kovariate (eindimensional)
$\boldsymbol{Z}$	Kovariatenvektor
U	Personen-Variable
R	Response-Indikator
τ_x	True-Outcome-Variable
$\delta_{xx'}$	True-Effect-Variable
C_X	Umfassende Kovariate
δ_{adj}	Adjustierte Effektvariable

Indizes

$J+1$	Anzahl der Treatment-Bedingungen; $x = 0, 1, \ldots, J$
Q	Anzahl der Kovariaten; $q = 1, \ldots, Q$
M	Anzahl der Wertekombinationen des Kovariatenvektors $\boldsymbol{Z}$; $m = 1, \ldots, M$
T	Indexmenge; $T = \{1, 2, \ldots, n\}$ mit $n \in \mathbb{N}$, wobei n die Anzahl der σ-Algebren in der Filtration $(\mathfrak{F}_t)_{t \in T}$ darstellt

Abkürzungen

$\tau_x \vdash X \mid \boldsymbol{Z},\ \forall x$	Kurzform von: $\forall x \in \Omega_X : \tau_x \vdash X \mid \boldsymbol{Z}$, $\boldsymbol{Z}$-bedingte regressive Unabhängigkeit der True-Outcome-Variable τ_x von der Treatment-Variable X
$X \perp\!\!\!\perp \tau_x \mid \boldsymbol{Z},\ \forall x$	Kurzform von: $\forall x \in \Omega_X : X \perp\!\!\!\perp \tau_x \mid \boldsymbol{Z}$, $\boldsymbol{Z}$-bedingte stochastische Unabhängigkeit der Treatment-Variable X und der True-Outcome-Variable τ_x
$X \perp\!\!\!\perp C_X \mid \boldsymbol{Z}$	$\boldsymbol{Z}$-bedingte stochastische Unabhängigkeit der Treatment-Variable X und der umfassenden Kovariate C_X
$Y \vdash C_X \mid X, \boldsymbol{Z}$	$\boldsymbol{Z}$-bedingte regressive Unabhängigkeit der Outcome-Variable Y von der umfassenden Kovariate C_X gegeben X

B Vergleichsarbeiten und weitere Formen der Evaluation

Vergleichsarbeiten sind mittlerweile zu einem etablierten Werkzeug der Qualitätssicherung im deutschen Bildungswesen avanciert. Schon frühere Recherchen (z. B. Hovestadt & Kessler, 2005; van Ackeren & Bellenberg, 2004) ergaben, dass die Bundesländer Vergleichsarbeiten durchführen. Um diesbezüglich die aktuellen Entwicklungen abbilden zu können, wurde im Rahmen des BMBF-Projektes *Faire Vergleiche* erhoben, *was* und *wie* die einzelnen Bundesländer zur evidenzbasierten Qualitätssicherung im Bildungswesen beitragen.

Mit Hilfe einer Literaturrecherche sowohl auf den Webseiten der Bundesländer als auch in Datenbanken (z. B. PsycINFO) wurden relevante Informationen gesammelt. Der Schwerpunkt der Recherchearbeit lag auf den Vergleichsarbeiten; insbesondere auf der Art und Weise der Berechnung *fairer Vergleiche*. Zudem wurden Informationen zu weiteren Evaluationsformen gesammelt. Waren die Informationen nicht in wissenschaftlichen Veröffentlichungen zu finden, wurde mit den entsprechenden Institutionen (Landesinstitute bzw. zuständige Ministerien der Länder) per Mail und/oder telefonisch Kontakt aufgenommen, um die fehlenden Informationen in Erfahrung zu bringen.

Auf Basis der Rechercheergebnisse wurde eine Klassifikation[1] der verschiedenen Evaluationsformen erarbeitet (vgl. Tabelle B.1 und B.2). Diese Kategorisierung erlaubt – neben einer Übersicht der Qualitätssicherungs- und Qualitätsentwicklungsmaßnahmen im Bildungskontext – die Abgrenzung von Vergleichsarbeiten. Abgrenzungsmerkmale sind u. a. die Ziele einer Schulleistungsuntersuchung (z. B. individuelle Lernförderung) und der Fokus der Evaluationsintention, d. h. ob es sich um eine formative oder summative Evaluation handelt.

[1] Diese Klassifikation erhebt keinen Anspruch auf Vollständigkeit.

Tabelle B.1: Klassifikation von Schulleistungsuntersuchungen und anderen Evaluationsformen im Bildungskontext

Bezeichnung	**Kurzbeschreibung**	**(Haupt-) Ziele**	**Evaluations-form (Fokus)**	**Beispiele**
(1) Vergleichsarbeiten	– standardisierte Schulleistungstests in Anlehnung an die Bildungsstandards – Testentwicklung zentral (IQB) auf Basis psychometrischer Testgütekriterien; Auswertung länderintern (z. B. Landesinstitute) – Rückmeldung an Schulen, Klassen & Schüler – Evaluation im Verlauf der Schullaufbahn	Unterrichts- & Schulentwicklung	formativ & summativ	Kompetenztests in Thüringen
(2) Parallelarbeiten	– Parallelklassen (i. d. R. innerhalb einer Schule) erhalten identische Testaufgaben – Testentwicklung, -auswertung & -benotung durch Lehrkräfte	Unterrichts- & Schulentwicklung	formativ & summativ	Parallelarbeiten in Bremen
(3) Zentrale Abschluss-prüfungen	– länderinterne, zentrale Erfassung der Abschlussleistungen (am Ende der Pflicht-schulzeit bzw. der Sekundarstufe II); schulform- & fächerspezifisch – Durchführung zum gleichen Zeitpunkt mit standardisierten Testaufgaben – Evaluation am Ende der Schullaufbahn	Zertifizierung	summativ	MSA[a] in Berlin
(4) Diagnose- & Förderinstrumente	– Förderdiagnosen (Feststellung von Förder-bedarf) & ggf. individualisiertes Fördermaterial – i. d. R. keine sozialen, sondern kriteriale Vergleiche mit Cutoff-Kriterium	Individuelle Lernförderung	formativ	LAL 7[b] in Berlin

Anmerkungen. [a] MSA = Mittlerer Schulabschluss, [b] LAL 7 = Ermittlung der Lernausgangslagen in Klassenstufe 7

Tabelle B.2: Klassifikation von Schulleistungsuntersuchungen und anderen Evaluationsformen im Bildungskontext (Fortsetzung)

Bezeichnung	**Kurzbeschreibung**	**(Haupt-) Ziele**	**Evaluations-form (Fokus)**	**Beispiele**
(5) Schulinspektionen	– Externe Schulevaluation *und* Evaluations-systeme mit einer Kopplung von interner & externer Schulevaluation – Fokus liegt auf Schule als System (Organisationsentwicklung)	Schulentwicklung	formativ	EVAS[a] in Thüringen
(6) Entwicklungs-programme	– (Weiter-) Entwicklung didaktischer Konzepte, die formativ evaluiert werden – umfasst u. a. Lehrertrainings & lehrbezogene Feedbacksysteme	Unterrichts-entwicklung	formativ	E.U.LE.[b] in Thüringen
(7) Nationale Schulleistungs-untersuchungen	– wissenschaftliche Untersuchung von Lernständen und den Bedingungen von Lehre & Lernen – aggregierte Rückmeldung i. d. R. auf Systemebene (z. B. auf Ebene einzelner Bundesländer)	Unterrichts- & Schulentwicklung	summativ	TOSCA[c], KMK-Länder-vergleiche
(8) Internationale Schulleistungs-untersuchungen	– wissenschaftliche Untersuchung von Lernständen und den Bedingungen von Lehre & Lernen – aggregierte Rückmeldung i. d. R. auf Systemebene (z. B. auf Ebene einzelner Nationen)	Unterrichts- & Schulentwicklung	summativ	TIMSS, PIRLS/IGLU, PISA

Anmerkungen. [a] EVAS = Eigenverantwortliche Schule, [b] E.U.LE. = Entwicklungsprogramm für Unterricht und Lernqualität, [c] TOSCA = Transformation des Sekundarschulsystems und akademische Karrieren

C Contextual Models

In Kapitel 4 in Abschnitt 4.2.2 wurde ein Contextual Model (CM) dargestellt, bei dem die Zentrierung der Ebene-1-Prädiktorvariable am Gesamtmittelwert vorgenommen wurde (*grand-mean centering*). Alternativ kann die Zentrierung auch am Gruppenmittelwert erfolgen (*group-mean centering*). Dies hat zur Folge, dass sich die Bedeutung und Interpretation der Regressionskoeffizienten sowie die Schätzung des Kontexteffekts verändern. Die Definition und Berechnung von Kontexteffekten in der Literatur basiert nicht selten auch auf dieser Zentrierung (vgl. z. B. Enders & Tofighi, 2007; Lüdtke et al., 2008). Um Konfusionen mit entsprechender Literatur und der in Abschnitt 4.2.2 gewählten Darstellung zu vermeiden, wird zum Zwecke der Vollständigkeit nachfolgend auch dieses alternative CM dargestellt.

Wir betrachten erneut folgende exemplarische Analyse hierarchisch-strukturierter Daten mit zwei Ebenen: Schüler (Ebene 1), die verschiedenen Schulen (Ebene 2) angehören. Eine abhängige Variable Y, deren Werte die Mathematikleistungen der Schüler sind, soll mittels einer Ebene-1-Variable Z^*_{E1}, dem SES auf Schülerebene, sowie deren Aggregat auf Schulebene Z_{E2} vorhergesagt werden. Dabei sei $Z_{E2} = \bar{Z}_{\bullet j}$, d. h. die Ebene-2-Variable ist der durchschnittlicher SES einer Schule j. Dann lautet die Modellgleichung auf Schülerebene (Ebene-1-Modell) – entsprechend der Notation nach Raudenbush und Bryk (2002) – wie folgt:

$$Y_{ij} = \beta^*_{0j} + \beta^*_{1j} \cdot (Z_{ij} - \bar{Z}_{\bullet j}) + r^*_{ij} \,, \tag{C.1}$$

wobei Y_{ij} die Mathematikleistung eines Schülers i in Schule j ist, die durch die zentrierte SES-Variable vorhergesagt wird. Im Gegensatz zu dem CM in Kapitel 4 (Abschnitt 4.2.2) ist die SES-Variable nun um den Gruppenmittelwert zentriert, d. h. $Z^*_{E1} = Z_{ij} - \bar{Z}_{\bullet j}$ (*group-mean centering*; Zentrierung um den Gruppenmittelwert). Weiterhin sind β^*_{0j} das Interzept, β^*_{1j} der Anstieg und r^*_{ij} das Ebene-1-Residuum. Die Modellgleichungen auf Ebene 2 eines *Random-Intercept-Modells* im Rahmen einer Kontextanalyse lauten

dann folgendermaßen:

$$\beta_{0j} = \gamma^*_{00} + \gamma^*_{01} \cdot \bar{Z}_{\bullet j} + u^*_{0j} \,, \tag{C.2}$$

$$\beta_{1j} = \gamma^*_{10} \,, \tag{C.3}$$

wobei γ^* die festen Effekte und u^*_{0j} die zufälligen Effekte sind. Durch Einsetzen der Ebene-2-Modellgleichungen in das Ebene-1-Modell erhält man schließlich das gemischte Modell (*linear mixed effect notation*; vgl. McCulloch & Searle, 2001):

$$Y_{ij} = \gamma^*_{00} + \gamma^*_{10} \cdot (Z_{ij} - \bar{Z}_{\bullet j}) + \gamma^*_{01} \cdot \bar{Z}_{\bullet j} + u^*_{0j} + r^*_{ij} \,. \tag{C.4}$$

Da Ebene-1-Prädiktoren, die um den Gruppenmittelwert zentriert sind, mit Ebene-2-Prädiktoren unkorreliert sind, enthält Gleichung C.4 zwei orthogonale Prädiktoren: γ^*_{10} bzw. γ^*_{01} quantifizieren jeweils den separaten (und nicht den partiellen) Zusammenhang zwischen der Mathematikleistung und dem SES auf Ebene 1 respektive auf Ebene 2. γ^*_{10} wird auch als *within-group regression coefficient* und γ^*_{01} als *between-group regression coefficient* bezeichnet (Cronbach, 1976). Ein Kontexteffekt liegt dann vor, wenn sich beide Koeffizienten statistisch unterscheiden (vgl. z. B. Enders & Tofighi, 2007), d. h. wenn gilt:

$$\gamma^*_{01} - \gamma^*_{10} \neq 0 \,. \tag{C.5}$$

Der Kontexteffekt $\gamma^*_{context}$ berechnet sich dann aus der Differenz zwischen dem *between-group*- und dem *within-group*- Regressionskoeffizienten:

$$\gamma^*_{context} = \gamma^*_{01} - \gamma^*_{10} \,. \tag{C.6}$$

Der Parameter $\gamma^*_{context}$ aus dem CM mit Zentrierung um den Gruppenmittelwert ist äquivalent mit dem Parameter γ_{01} des CM mit Zentrierung um den Gesamtmittelwert (vgl. Gleichung 4.4 in Kapitel 4), d. h. es gilt $\gamma^*_{context} = \gamma_{01}$.

D Struktur fehlender Werte

Aus der Annahme *vollständig zufällig* fehlender Werte (MCAR) folgt, dass es keinerlei Unterschiede in der Verteilung einer Variablen zwischen Schülern mit Missings und Schülern ohne Missings auf allen anderen Variablen gibt. Sobald dies für mindestens eine der Variablen nicht zutrifft, kann nicht mehr von MCAR für den Datensatz ausgegangen werden. Für die Mathematikleistung in Klassenstufe 8 wurde dies bereits in Kapitel 7 (vgl. Abschnitt 7.1.2) falsifiziert. Zum Zwecke der Vollständigkeit werden nachfolgend die entsprechenden Analysen für die zweite abhängige Variable – die Deutschleistung in Klassenstufe 8 (DK8) – dargestellt.

In Abbildung D.1 werden parallele Boxplots für die Verteilung der beobachteten Deutschleistungsscores in Klassenstufe 8 (DK8) jeweils in Abhängigkeit von der Missing-Struktur auf allen anderen Variablen dargestellt. Auf der linken Seite von Abbildung D.1 ist die Verteilung der Variable DK8 als Boxplot (weiß) abgebildet. Diese Verteilung wird (rechts daneben) insgesamt elf Mal in jeweils zwei Gruppen betrachtet. Die Gruppen entsprechen dabei den Fällen mit beobachteten (blau) bzw. mit fehlenden (rot) Werten auf allen anderen Variablen des Datensatzes (MK8, DK6, MK6, DK3L, DK3S, MK3, BLSF.M, WDH, SES.M, SES.D und MUSPR). Die Variablen Schulart (SART.M und SART.D) und Geschlecht (SEX) sind nicht in der Grafik enthalten, da für diese keine fehlenden Werte vorliegen. Des Weiteren ist die Variable BLSF.D nicht aufgeführt, da aufgrund einer zu geringen Missing-Anzahl[1] kein Boxplot für diese Gruppe dargestellt werden kann. Der erste blaue Boxplot stellt die Verteilung von DK8 für alle Fälle des Datensatzes dar, für die gleichfalls eine Beobachtung auf der Variable der MK8 vorliegt. Entsprechend zeigt der erste rote Boxplot die Verteilung von DK8 für die Fälle des Datensatzes, für welche die Werte der Variable MK8 fehlen. Hier wird ein geringer Unterschied bezüglich des Medians zwischen den beiden Gruppen sichtbar. Dieser Unterschied bezüglich des Medians von DK8 ist noch deutlicher zwischen

[1] Hier ist nicht die Gesamtanzahl fehlender Werte auf dieser Variable entscheidend, sondern die Anzahl der Missings auf dieser Variable für die Fälle mit vollständigen Werten auf DK8.

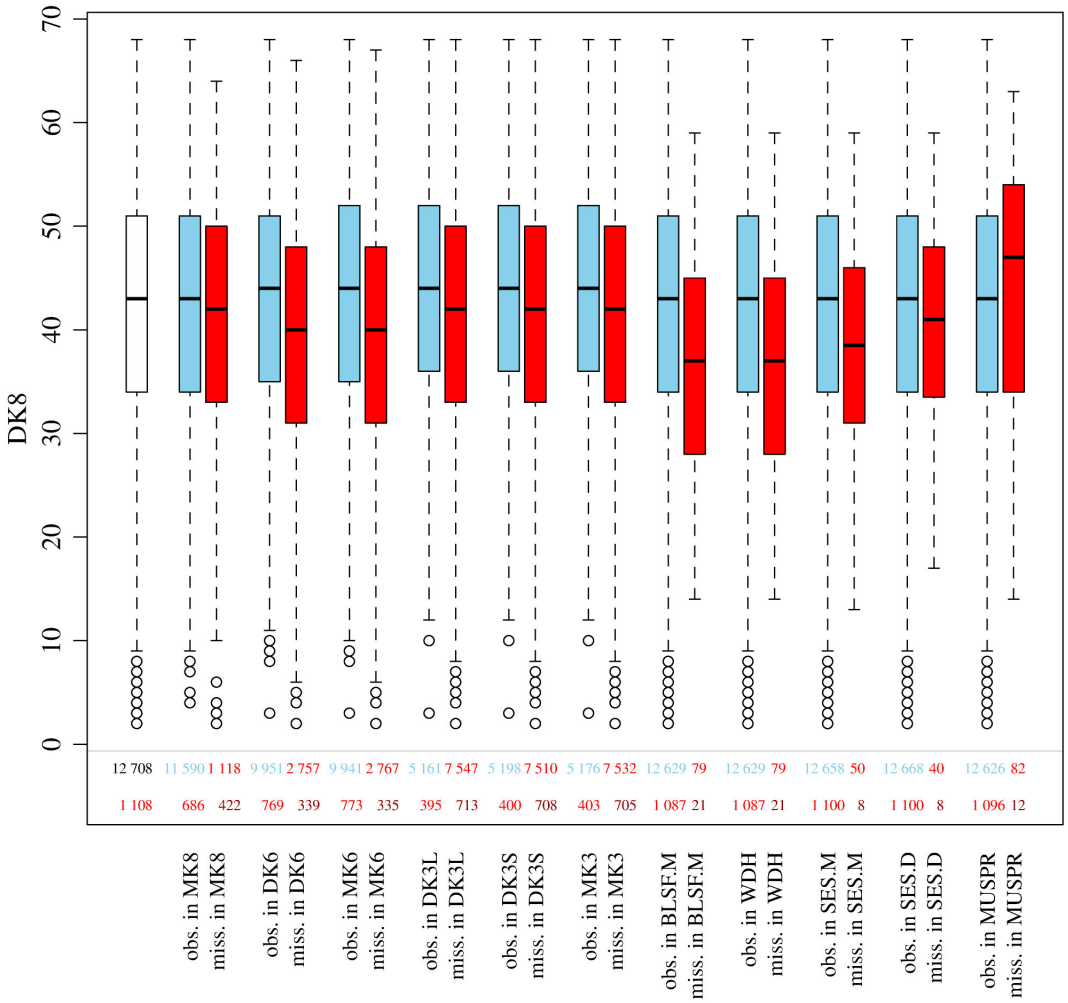

Abbildung D.1: Der weiße Boxplot (links) zeigt die Verteilung der beobachteten Deutschleistungsscores in Klassenstufe 8 (DK8). Rechts daneben: Parallele Boxplots für die Verteilung von DK8 in Abhängigkeit von der Missing-Struktur bezüglich der Variablen MK8, DK6, MK6, DK3L, DK3S, MK3, BLSF.M, WDH, SES.M, SES.D und MUSPR. Die Verteilung von DK8 wird hier in je zwei Gruppen dargestellt; getrennt nach dem Fehlen (rot = *missing*) und Nicht-Fehlen (blau = *observed*) auf den anderen Variablen des Datensatzes. Unterhalb der Boxplots sind die absoluten Häufigkeiten der beobachteten bzw. fehlenden Werte abgetragen.

Tabelle D.1: Mittelwerte, t-Test und Effektstärken mit der Deutschleistung in Klassenstufe 8 (DK8) als abhängige Variable und den Indikatorvariablen für fehlende Werte (Response-Indikatoren) als unabhängige Variablen

Response-Indikator[a]	M_{DK8} für $R_{[.]}$=1	M_{DK8} für $R_{[.]}$=0	t-Wert	(p-Wert)	d
R_{MK8}	42.33	41.08	2.78	(.006)	0.11
R_{DK6}	42.98	39.49	13.06	(.000)	0.31
R_{MK6}	43.02	39.35	13.84	(.000)	0.33
R_{DK3L}	43.29	41.53	8.34	(.000)	0.15
R_{DK3S}	43.23	41.57	7.91	(.000)	0.15
R_{MK3}	43.29	41.53	8.38	(.000)	0.16
$R_{BLSF.M}$	42.28	36.91	3.39	(.001)	0.47
R_{WDH}	42.28	36.91	3.39	(.001)	0.47
$R_{SES.M}$	42.27	38.26	2.40	(.021)	0.35
$R_{SES.D}$	42.26	40.16	1.11	(.276)	0.19
R_{MUSPR}	42.24	44.20	-1.24	(.218)	-0.17

Anmerkungen. [a] Indikatorvariable $R_{[.]}$, die mit dem Wert 1 anzeigt, dass für die entsprechende Variable [.] Beobachtungen vorliegen. Der Wert 0 zeigt an, dass keine Beobachtungen (Missings) für diese Variable [.] vorliegen (vgl. Gleichung 6.11).
[a] Als Effektstärkemaß wird Cohen's d (Cohen, 1988) verwendet.

den beiden Gruppen, die sich durch das Vorhandensein bzw. das Fehlen der Werte auf der Variable DK6 auszeichnen. Der Median der Gruppe, in der Beobachtungen für die Variable DK6 vorliegen, ist dabei größer als der in der Gruppe mit Missings auf dieser Variable. Dieser Befund zeigt sich gleichfalls für die restlichen Variablen, wobei lediglich bezüglich der Variablen Muttersprache (MUSPR) ein geringerer Medianwert in der Gruppe ohne Missings (im Vergleich zur Gruppe mit Missings) auf der jeweiligen Variable beobachtet wird.

Wie bereits in Kapitel 7 (vgl. Abschnitt 7.1.2) für die Variable MK8 wird die grafische Diagnose der Struktur fehlender Werte auch für die Variable DK8 gemäß den Empfehlungen von Peugh und Endes (2004) ergänzt. Dazu werden wiederum die Response-Indikatoren für jede der Variablen verwendet, um die Mittelwertsunterschiede zwischen der Gruppe mit beobachteten und der Gruppe mit fehlenden Werten mittels t-Tests zu betrachten. Tabelle D.1 enthält Mittelwerte, t-Werte und Effektstärken der Mittelwerts-

vergleiche bezüglich der Deutschleistung in Klassenstufe 8 (DK8). Hierbei wurden die Response-Indikatoren der elf Variablen betrachtet, die auch der grafischen Diagnose zugrunde lagen (vgl. Abbildung D.1). Neun der elf Mittelwertsunterschiede werden auf einem Signifikanzniveau von .05 signifikant. So ist bspw. der Mittelwert von DK8 der Schüler, die einen beobachteten Wert auf DK6 haben, signifikant höher als der durchschnittliche DK8-Score aller Schüler, die auf DK6 einen fehlenden Wert aufweisen ($t = 13.06$, $p = .000$). Die Effektstärke liegt bei $d = 0.31$. Dies widerspricht der Annahme, dass die Wahrscheinlichkeit fehlender Werte einer Variablen unabhängig ist von anderen Variablen im Datensatz. Die Beträge der Effektstärken für sämtliche der elf Vergleiche reichen von 0.11 bis 0.47. Die durchschnittliche Effektstärke beträgt $d = 0.26$. Diese Befunde entsprechen den Ergebnissen der grafischen Diagnose. Zudem ergänzen diese Ergebnisse die Befunde aus der Missinganalyse bezüglich der Mathematikleistung MK8 (vgl. Kapitel 7). Die Wahrscheinlichkeit fehlender Werte einer Variablen ist nicht unabhängig anderen Variablen im Datensatz. Die Annahme, dass die fehlenden Werte MCAR (d. h. *vollständig zufällig*) sind, wird somit verworfen.

E Standardfehler der Effektschätzungen

In Kapitel 6 wurde dargestellt, dass die Berechnung des adjustierten klassenspezifischen Effekts $E(\delta_{adj} \mid X{=}x)$ schrittweise erfolgt. Daher wurden auch die zugehörigen Standardfehler in einem iterativen Verfahren ermittelt. Dies wird im nachfolgenden Abschnitt gezeigt.

Vorgehen bei der Datenanalyse im Projekt *Kompetenztest.de*. Die Analyseschritte zur Berechnung des adjustierten klassenspezifischen Effekts $E(\delta_{adj} \mid X{=}x)$ bzw. seines Schätzers $\bar{\delta}_{adj;x}$ umfassen (1) die Berechnung des Klassenmittelwertes der Testwerte, (2) die Berechnung des adjustierten Referenzwertes einer Klasse und schließlich (3) die Berechnung des adjustierten klassenspezifischen Effektmaßes. Zum Zwecke der Verständlichkeit seien hier zunächst diese drei Analyseschritte zusammenfassend in Form mathematischer Gleichungen dargestellt (vgl. Kapitel 6; Abschnitt 6.3.1):

(1) *Klassenmittelwert der Testwerte:*

$$\bar{Y}_x = \widehat{E}(Y \mid X{=}x) \tag{E.1}$$

(2) *Adjustierter Referenzwert einer Klasse:*

$$\begin{aligned} \bar{Y}_{adj;x} &= \widehat{E}[\, \widehat{E}(Y \mid \mathbf{Z}) \mid X{=}x \,] \\ &= N_x^{-1} \cdot \sum_{i=1}^{N_x} \widehat{E}(Y \mid \mathbf{Z}{=}\boldsymbol{z}_i) \\ &= N_x^{-1} \cdot \sum_{i=1}^{N_x} \hat{\theta}_{\boldsymbol{z}_i}\,, \end{aligned} \tag{E.2}$$

wobei $\widehat{E}(Y \mid \boldsymbol{Z}{=}z_i) = \hat{\theta}_{z_i}$ die adjustierten individuellen (d. h. schülerspezifischen) Testwerte sind.

(3) *Adjustierter klassenspezifischer Effekt:*

$$\begin{aligned} \bar{\delta}_{adj;x} &= \widehat{E}(Y \mid X{=}x) - \widehat{E}\big[\, \widehat{E}(Y \mid \boldsymbol{Z}) \mid X{=}x \,\big] \\ &= \bar{Y}_x - \bar{Y}_{adj;x} \\ &= \widehat{E}(\delta_{adj} \mid X{=}x) \end{aligned} \tag{E.3}$$

Standardfehler der adjustierten klassenspezifischen Effekte. Der Standardfehler des klassenspezifischen Effektschätzers $SE(\bar{\delta}_{adj;x})$ ist dann:

$$\begin{aligned} SE(\bar{\delta}_{adj;x}) &= \sqrt{Var(\bar{Y}_x - \bar{Y}_{adj;x})} \\ &= \sqrt{Var(\bar{Y}_x) + Var(\bar{Y}_{adj;x}) - 2 \cdot Cov\,(\bar{Y}_x, \bar{Y}_{adj;x})} \\ &= \sqrt{Var(\bar{Y}_x) + Var(\bar{Y}_{adj;x})} && [Cov\,(\bar{Y}_x, \bar{Y}_{adj;x}) = 0] \\ &= \sqrt{SE(\bar{Y}_x)^2 + SE(\bar{Y}_{adj;x})^2} \end{aligned} \tag{E.4}$$

Der Standardfehler des Effektschätzers $SE(\bar{\delta}_{adj;x})$ ist somit die Wurzel aus der Summe der quadrierten Standardfehler des Klassenmittelwertes und des adjustierten Klassenmittelwertes. Hier wird die Annahme gemacht, dass die Schätzer $\bar{Y}_x$ und $\bar{Y}_{adj;x}$ korrelativ unabhängig sind[1], d. h., dass gilt: $Cov\,(\bar{Y}_x, \bar{Y}_{adj;x}) = 0$ (vgl. dritte Zeile von Gleichung E.4). Da die Schätzung sowohl von $\bar{Y}_x$ als auch von $\bar{Y}_{adj;x}$ separat erfolgt, werden die Summanden der vierten Zeile in Gleichung E.4 separat geschätzt.

In einem ersten Schritt wird der Standardfehler der Klassenmittelwerte $SE(\bar{Y}_x)$ wie folgt berechnet:

$$\begin{aligned} SE(\bar{Y}_x) &= \sqrt{Var(\bar{Y}_x)} \\ &= \sqrt{N_x^{-1}} \cdot SD(Y \mid X{=}x) \end{aligned} \tag{E.5}$$

[1] Die Plausibilität der Annahme korrelativer Unabhängigkeit der Schätzer $\bar{Y}_x$ und $\bar{Y}_{adj;x}$ im vorliegenden Anwendungskontext ist fraglich. Inhaltlich plausibel ist auch eine positive Korrelation, was zu einer Überschätzung der Standardfehler in Gleichung E.4 führen würde. Aus pragmatischen Gründen (sowie aus Mangel an Alternativen) wähle ich jedoch dennoch dieses Vorgehen zur Berechnung der Standardfehler.

In einem zweiten Schritt erfolgt die Berechnung des Standardfehlers der adjustierten Referenzwerte $SE(\bar{Y}_{adj;x})$:

$$
\begin{aligned}
SE(\bar{Y}_{adj;x}) &= \sqrt{Var(\bar{Y}_{adj;x})} \\
&= \sqrt{Var(N_x^{-1} \cdot \sum_{i=1}^{N_x} \hat{\theta}_{z_i})} \\
&= \sqrt{N_x^{-2} \cdot \sum_{i=1}^{N_x} Var(\hat{\theta}_{z_i})} \qquad [Cov\,(\hat{\theta}_{z_i}, \hat{\theta}_{z_j}) = 0,\ i \neq j] \\
&= N_x^{-1} \cdot \sqrt{\sum_{i=1}^{N_x} Var(\hat{\theta}_{z_i})} \\
&= N_x^{-1} \cdot \sqrt{\sum_{i=1}^{N_x} SE(\hat{\theta}_{z_i})^2}\ ,
\end{aligned}
\tag{E.6}
$$

wobei

$$
SE(\hat{\theta}_{z_i}) = \sqrt{z_i^T VCOV(\beta) z_i}\ . \tag{E.7}
$$

Die in Gleichung E.5 und Gleichung E.6 berechneten Maße werden schließlich gemäß Gleichung E.4 zum Standardfehler des adjustierten klassenspezifischen Effektschätzers $SE(\bar{\delta}_{adj;x})$ kombiniert. In konkreten empirischen Anwendungen erhält man nicht den (theoretischen) Standardfehler des Effektschätzers $SE(\bar{\delta}_{adj;x})$, sondern wiederum eine Schätzung dieses Parameters, da auch die in Gleichung E.4 bis E.6 angegebenen Varianzen der Parameterschätzer geschätzt werden müssen. In einer konkreten empirischen Anwendung erhält man daher den geschätzten Standardfehler $\widehat{SE}(\bar{\delta}_{adj;x})$.

Kombination der Standardfehler über die Imputationen. Auf die soeben dargestellte Weise wurden die Standardfehler für jeden der m=5 imputierten Datensätze berechnet und anschließend nach dem von Rubin (1987) beschriebenen Verfahren zu einer Gesamtschätzung kombiniert (vgl. Peugh & Enders, 2004; Rubin, 1987). Bei der Berechnung eines kombinierten Standardfehlers SE_{pooled} – dem sog. *pooled standard error* – müssen zwei Variationsquellen der Schätzung berücksichtigt werden: (a) die jeweilige Varianz innerhalb eines imputierten Datensatzes (*within-imputation variance*) und (b) die Varianz der Parameterschätzungen zwischen den m=5 imputierten Daten-

sätzen (*between-imputation variance*). Dementsprechend erfolgt auch die Kombination der Standardfehler über die imputierten Datensätze hinweg in mehreren Analyseschritten, die nachfolgend dargestellt werden (Peugh & Enders, 2004).

(1) *Berechnung der within-imputation variance $\bar{U}$:*
Die *within-imputation variance* ist das arithmetische Mittel der quadrierten Standardfehlerschätzungen der m=5 Analysen, die für jeden der imputierten Datensatz durchgeführt wurden:

$$\bar{U} = \frac{1}{m} \sum_{k=1}^{m} \hat{U}_k \ . \tag{E.8}$$

Im vorliegenden Anwendungsfall gilt: $\hat{U}_k = \widehat{SE}(\bar{\delta}_{adj;x})_k^2$, d. h. $\hat{U}_k$ ist das Quadrat des geschätzten Standardfehlers aus dem $k-$ten imputierten Datensatz.

(2) *Berechnung der between-imputation variance B:*
Die *between-imputation variance* quantifiziert die Variabilität der m Parameterschätzungen:

$$B = \frac{1}{m} \sum_{k=1}^{m} (\hat{Q}_k - \bar{Q})^2 \ , \tag{E.9}$$

wobei

$$\bar{Q} = \frac{1}{m} \sum_{k=1}^{m} \hat{Q}_k \ . \tag{E.10}$$

Hier ist $\hat{Q}_k$ die Schätzung des klassenspezifischen Effekts des $k-$ten imputierten Datensatzes.

(3) *Berechnung des kombinierten Standardfehlers SE_{pooled}:*
Die kombinierte bzw. gepoolte Varianz lässt sich schließlich aus beiden Varianzen $\bar{U}$ und B berechnen:

$$T = \bar{U} + (1 + \frac{1}{m}) \cdot B \ . \tag{E.11}$$

Die positive Quadratwurzel aus T ist schließlich der über die m Imputationen kombinierte Standardfehler, d. h. $\sqrt{T} = SE_{\text{pooled}}$.

F Sensitivität der adjustierten Effektschätzungen: Supplement

Um die Sensitivität der adjustierten klassenspezifischen Effektschätzungen gegenüber der Kovariaten- und Modellselektion, wurden in Kapitel 7 u. a. die Veränderungen des Quintil-Rankings der klassenspezifischen Effektschätzungen infolge des Wechsels von einem zu einem anderen Modell betrachtet. Dafür wurden *Transitionsmatrizen* verwendet. Aus Gründen der Übersichtlichkeit wurden in Kapitel 7 nur die zur Beurteilung der Plausibilität der postulierten Hypothesen notwendigen Transitionsmatrizen dargestellt. Nachfolgend sollen zum Zwecke der Vollständigkeit auch die weiteren Transitionsmatrizen abgebildet werden – zunächst für den Fachbereich Mathematik und anschließend für Deutsch. Die Ergebnisse sind konkordant zu dem in Kapitel 7 berichteten Ergebnismuster.

Das im Rahmen dieser Arbeit verwendete Design des Modellvergleichs erlaubt – bei jeweils konstanter Parametrisierung der Modelle – insgesamt jeweils drei Vergleiche beim Wechsel vom CAM zum VAM. Diese sind in Tabelle F.1 für Mathematik und in Tabelle F.4 für Deutsch dargestellt. Dies trifft gleichermaßen für den Modellwechsel vom VAM zum CVA zu. Die entsprechenden Transitionsmatrizen finden sich in Tabelle F.2 (Mathematik) und Tabelle F.5 (Deutsch). In allen vier Tabellen sind auf der jeweils linken Seite die Transitionsmatrizen von Modellen mit bedingt linearer Parametrisierung (inkl. Interaktionen) und auf der jeweils rechten Seite von Modellen mit linearer Parametrisierung dargestellt. Schließlich zeigen die Tabellen F.3 und F.6 die Transitionsmatrizen infolge des Wechsels der Parametrisierung – jeweils links für die VAM bzw. rechts für die CVA.

Tabelle F.1: Veränderungen des Quintil-Rankings der adjustierten Effektschätzungen im Fach Mathematik (MK8) infolge der zusätzlichen Berücksichtigung des fachspezifischen Vorwissens: CAM *versus* VAM

Bedingt lineare Parametrisierung							**Lineare Parametrisierung**						
		Modell 2							Modell 9				
	Quintil	1	2	3	4	5		Quintil	1	2	3	4	5
Modell 1	1	86.90	13.10	0	0	0	Modell 8	1	86.21	13.79	0	0	0
	2	11.72	66.90	20.69	0.69	0		2	13.79	66.90	19.31	0	0
	3	1.39	19.44	61.11	18.06	0		3	0	19.44	60.42	19.44	0.69
	4	0	0.69	17.24	68.97	13.10		4	0	0	20.00	66.90	13.10
	5	0	0	0.69	12.41	86.90		5	0	0	0	13.79	86.21
		Modell 3							Modell 10				
Modell 1	1	75.17	22.07	1.38	1.38	0	Modell 8	1	73.79	24.14	2.07	0	0
	2	19.31	42.76	32.41	4.83	0.69		2	24.14	43.45	23.45	7.59	1.38
	3	4.17	26.39	36.81	27.08	5.56		3	1.39	25.69	42.36	27.08	3.47
	4	1.38	8.28	24.83	44.83	20.69		4	0.69	5.52	25.52	44.14	24.14
	5	0	0.69	4.14	22.07	73.10		5	0	1.38	6.21	21.38	71.03
		Modell 4							Modell 11				
Modell 1	1	73.10	22.07	4.14	0.69	0	Modell 8	1	71.72	25.52	2.76	0	0
	2	21.38	39.31	31.03	7.59	0.69		2	25.52	40.69	23.45	8.28	2.07
	3	4.17	27.78	33.33	29.86	4.86		3	2.08	25.69	40.28	26.39	5.56
	4	1.38	10.34	26.90	40.69	20.69		4	0.69	6.90	25.52	44.14	22.76
	5	0	0.69	4.14	21.38	73.79		5	0	1.38	7.59	21.38	69.66

Anmerkungen. Die Werte in den Zellen der Tabelle sind Prozent der Klassen pro Zeile, deren Effektschätzung im entsprechenden Quintilsabschnitt liegt. Die zeilenweise Summe beträgt somit jeweils 100%. Anzahl der Klassen: $N = 724$. Anzahl der Klassen pro Zeile bzw. pro Spalte: $n_1 = 145$, $n_2 = 145$, $n_3 = 144$, $n_4 = 145$, $n_5 = 145$.

Tabelle F.2: Veränderungen des Quintil-Rankings der adjustierten Effektschätzungen im Fach Mathematik (MK8) infolge der zusätzlichen Berücksichtigung der leistungsmäßigen Klassenkomposition: VAM *versus* CVA

Bedingt lineare Parametrisierung							**Lineare Parametrisierung**						
		Modell 5							Modell 12				
	Quintil	1	2	3	4	5		Quintil	1	2	3	4	5
Modell 2	1	84.14	12.41	3.45	0	0	Modell 9	1	85.52	14.48	0	0	0
	2	15.86	62.76	20.00	1.38	0		2	14.48	63.45	22.07	0	0
	3	0	22.22	60.42	17.36	0		3	0	22.22	54.86	22.22	0.69
	4	0	2.76	13.79	70.34	13.10		4	0	0	22.07	64.83	13.10
	5	0	0	2.07	11.03	86.90		5	0	0	0.69	13.10	86.21
		Modell 6							Modell 13				
Modell 3	1	79.31	15.86	4.83	0	0	Modell 10	1	82.07	17.24	0.69	0	0
	2	20.00	57.93	19.31	2.76	0		2	15.86	53.79	28.28	2.07	0
	3	0.69	22.22	51.39	25.00	0.69		3	2.08	27.78	47.92	21.53	0.69
	4	0	2.76	21.38	59.31	16.55		4	0	1.38	21.38	55.86	21.38
	5	0	1.38	2.76	13.10	82.76		5	0	0	1.38	20.69	77.93
		Modell 7							Modell 14				
Modell 4	1	75.86	19.31	4.14	0.69	0	Modell 11	1	82.76	16.55	0.69	0	0
	2	22.07	56.55	17.93	3.45	0		2	16.55	54.48	24.83	4.14	0
	3	2.08	21.53	55.56	20.14	0.69		3	0.69	27.78	54.17	16.67	0.69
	4	0	2.07	19.31	63.45	15.17		4	0	1.38	20.00	62.07	16.55
	5	0	0.69	2.76	12.41	84.14		5	0	0	0	17.24	82.76

Anmerkungen. Die Werte in den Zellen der Tabelle sind Prozent der Klassen pro Zeile, deren Effektschätzung im entsprechenden Quintilsabschnitt liegt. Die zeilenweise Summe beträgt somit jeweils 100%. Anzahl der Klassen: $N = 724$. Anzahl der Klassen pro Zeile bzw. pro Spalte: $n_1 = 145$, $n_2 = 145$, $n_3 = 144$, $n_4 = 145$, $n_5 = 145$.

Tabelle F.3: Veränderungen des Quintil-Rankings der adjustierten Effektschätzungen im Fach Mathematik (MK8): Bedingt lineare *versus* lineare Parametrisierung

Value-Added Model (VAM)							**Contextual Value-Added Model (CVA)**						
		Modell 9							Modell 12				
	Quintil	1	2	3	4	5		Quintil	1	2	3	4	5
Modell 2	1	77.93	20.00	1.38	0.69	0	Modell 5	1	74.48	22.76	2.76	0	0
	2	19.31	48.97	24.14	5.52	2.07		2	20.00	48.28	24.83	6.21	0.69
	3	2.78	24.31	43.75	26.39	2.78		3	4.86	27.08	38.19	21.53	8.33
	4	0	6.90	26.90	44.14	22.07		4	0.69	1.38	30.34	48.28	19.31
	5	0	0	3.45	23.45	73.10		5	0	0.69	3.45	24.14	71.72
		Modell 10							Modell 13				
Modell 3	1	80.00	17.24	2.76	0	0	Modell 6	1	79.31	16.55	4.14	0	0
	2	17.93	55.17	20.00	6.90	0		2	17.93	53.79	23.45	4.83	0
	3	2.08	22.22	47.92	24.31	3.47		3	2.78	25.00	40.28	25.69	6.25
	4	0	5.52	28.28	43.45	22.76		4	0	4.83	30.34	44.14	20.69
	5	0	0	0.69	25.52	73.79		5	0	0	1.38	25.52	73.10
		Modell 11							Modell 14				
Modell 4	1	78.62	17.24	4.14	0	0	Modell 7	1	77.24	19.31	3.45	0	0
	2	17.93	51.03	26.21	4.14	0.69		2	20.69	48.97	24.83	5.52	0
	3	3.47	26.39	44.44	22.92	2.78		3	2.08	23.61	41.67	24.31	8.33
	4	0	5.52	24.83	47.59	22.07		4	0	8.28	28.28	44.83	18.62
	5	0	0	0	25.52	74.48		5	0	0	1.38	25.52	73.10

Anmerkungen. Die Werte in den Zellen der Tabelle sind Prozent der Klassen pro Zeile, deren Effektschätzung im entsprechenden Quintilsabschnitt liegt. Die zeilenweise Summe beträgt somit jeweils 100%. Anzahl der Klassen: $N = 724$. Anzahl der Klassen pro Zeile bzw. pro Spalte: $n_1 = 145$, $n_2 = 145$, $n_3 = 144$, $n_4 = 145$, $n_5 = 145$.

Tabelle F.4: Veränderungen des Quintil-Rankings der adjustierten Effektschätzungen im Fach Deutsch (DK8) infolge der zusätzlichen Berücksichtigung des fachspezifischen Vorwissens: CAM *versus* VAM

Bedingt lineare Parametrisierung							**Lineare Parametrisierung**						
		Modell 2							Modell 9				
	Quintil	1	2	3	4	5		Quintil	1	2	3	4	5
Modell 1	1	83.69	14.18	2.13	0	0	Modell 8	1	81.56	17.73	0.71	0	0
	2	16.43	58.57	22.14	2.86	0		2	16.43	58.57	21.43	3.57	0
	3	0	25.00	53.57	17.86	3.57		3	2.14	22.14	51.43	22.86	1.43
	4	0	2.14	21.43	64.29	12.14		4	0	1.43	25.71	60.00	12.86
	5	0	0	0.71	14.89	84.40		5	0	0	0.71	13.48	85.82
		Modell 3							Modell 10				
Modell 1	1	78.72	13.48	4.26	1.42	2.13	Modell 8	1	75.18	16.31	4.96	1.42	2.13
	2	17.86	50.00	22.14	6.43	3.57		2	22.86	47.14	22.86	5.71	1.43
	3	2.86	30.00	41.43	22.14	3.57		3	1.43	30.00	40.71	20.00	7.86
	4	0.71	6.43	28.57	45.71	18.57		4	0.71	6.43	27.14	53.57	12.14
	5	0	0	3.55	24.11	72.34		5	0	0	4.26	19.15	76.60
		Modell 4							Modell 11				
Modell 1	1	77.30	12.77	6.38	2.13	1.42	Modell 8	1	71.63	18.44	6.38	2.84	0.71
	2	18.57	48.57	21.43	6.43	5.00		2	22.14	45.71	20.71	9.29	2.14
	3	3.57	27.14	40.71	22.14	6.43		3	5.71	27.86	35.71	23.57	7.14
	4	0.71	10.71	26.43	45.71	16.43		4	0.71	7.86	32.14	44.29	15.00
	5	0	0.71	4.96	23.40	70.92		5	0	0	4.96	19.86	75.18

Anmerkungen. Die Werte in den Zellen der Tabelle sind Prozent der Klassen pro Zeile, deren Effektschätzung im entsprechenden Quintilsabschnitt liegt. Die zeilenweise Summe beträgt somit jeweils 100%. Anzahl der Klassen: $N = 702$. Anzahl der Klassen pro Zeile bzw. pro Spalte: $n_1 = 141$, $n_2 = 140$, $n_3 = 140$, $n_4 = 140$, $n_5 = 141$.

Tabelle F.5: Veränderungen des Quintil-Rankings der adjustierten Effektschätzungen im Fach Deutsch (DK8) infolge der zusätzlichen Berücksichtigung der leistungsmäßigen Klassenkomposition: VAM *versus* CVA

Bedingt lineare Parametrisierung							**Lineare Parametrisierung**						
		Modell 5							Modell 12				
	Quintil	1	2	3	4	5		Quintil	1	2	3	4	5
Modell 2	1	85.82	11.35	2.84	0	0	Modell 9	1	89.36	10.64	0	0	0
	2	13.57	66.43	15.71	4.29	0		2	10.71	77.14	12.14	0	0
	3	0.71	21.43	58.57	17.86	1.43		3	0	12.14	66.43	20.71	0.71
	4	0	0.71	19.29	67.14	12.86		4	0	0	21.43	66.43	12.14
	5	0	0	3.55	10.64	85.82		5	0	0	0	12.77	87.23
		Modell 6							Modell 13				
Modell 3	1	78.72	18.44	2.84	0	0	Modell 10	1	97.16	2.84	0	0	0
	2	17.86	57.14	20.71	4.29	0		2	2.86	92.86	4.29	0	0
	3	2.14	19.29	60.00	18.57	0		3	0	4.29	88.57	7.14	0
	4	0.71	2.86	12.86	62.86	20.71		4	0	0	7.14	87.86	5.00
	5	0.71	2.13	3.55	14.18	79.43		5	0	0	0	4.96	95.04
		Modell 7							Modell 14				
Modell 4	1	81.56	14.18	3.55	0.71	0	Modell 11	1	90.07	9.93	0	0	0
	2	16.43	61.43	20.71	1.43	0		2	10.00	81.43	8.57	0	0
	3	2.14	20.71	53.57	22.86	0.71		3	0	8.57	80.00	11.43	0
	4	0	2.14	16.43	60.71	20.71		4	0	0	11.43	80.00	8.57
	5	0	1.42	5.67	14.18	78.72		5	0	0	0	8.51	91.49

Anmerkungen. Die Werte in den Zellen der Tabelle sind Prozent der Klassen pro Zeile, deren Effektschätzung im entsprechenden Quintilsabschnitt liegt. Die zeilenweise Summe beträgt somit jeweils 100%. Anzahl der Klassen: $N = 702$. Anzahl der Klassen pro Zeile bzw. pro Spalte: $n_1 = 141$, $n_2 = 140$, $n_3 = 140$, $n_4 = 140$, $n_5 = 141$.

Tabelle F.6: Veränderungen des Quintil-Rankings der adjustierten Effektschätzungen im Fach Deutsch (DK8): Bedingt lineare *versus* lineare Parametrisierung

Value-Added Model (VAM)							**Contextual Value-Added Model (CVA)**						
		Modell 9							Modell 12				
	Quintil	1	2	3	4	5		Quintil	1	2	3	4	5
Modell 2	1	73.05	24.82	1.42	0.71	0	Modell 5	1	78.01	20.57	0.71	0.71	0
	2	25.00	49.29	18.57	5.00	2.14		2	17.14	52.86	25.00	2.86	2.14
	3	2.14	23.57	53.57	17.14	3.57		3	4.29	21.43	50.00	14.29	10.00
	4	0	2.14	23.57	59.29	15.00		4	0.71	2.86	21.43	60.71	14.29
	5	0	0	2.84	17.73	79.43		5	0	2.13	2.84	21.28	73.76
		Modell 10							Modell 13				
Modell 3	1	77.30	18.44	4.26	0	0	Modell 6	1	75.89	19.15	3.55	0	1.42
	2	20.71	54.29	20.71	4.29	0		2	20.00	49.29	25.71	5.00	0
	3	2.14	23.57	50.71	22.14	1.43		3	3.57	25.00	42.14	21.43	7.86
	4	0	3.57	23.57	55.71	17.14		4	0.71	6.43	25.00	47.86	20.00
	5	0	0	0.71	17.73	81.56		5	0	0	3.55	25.53	70.92
		Modell 11							Modell 14				
Modell 4	1	78.01	20.57	1.42	0	0	Modell 7	1	73.05	24.82	0.71	1.42	0
	2	20.00	57.14	20.71	1.43	0.71		2	18.57	51.43	25.00	2.86	2.14
	3	2.14	20.71	51.43	23.57	2.14		3	7.86	20.71	42.14	16.43	12.86
	4	0	1.43	25.71	57.86	15.00		4	0.71	2.86	29.29	50.71	16.43
	5	0	0	0.71	17.02	82.27		5	0	0	2.84	28.37	68.79

Anmerkungen. Die Werte in den Zellen der Tabelle sind Prozent der Klassen pro Zeile, deren Effektschätzung im entsprechenden Quintilsabschnitt liegt. Die zeilenweise Summe beträgt somit jeweils 100%. Anzahl der Klassen: $N = 702$. Anzahl der Klassen pro Zeile bzw. pro Spalte: $n_1 = 141$, $n_2 = 140$, $n_3 = 140$, $n_4 = 140$, $n_5 = 141$.

Ehrenwörtliche Erklärung

Die Promotionsordnung der Fakultät für Sozial- und Verhaltenswissenschaften der Friedrich-Schiller-Universität in der geltenden Fassung ist mir bekannt.

Ich habe diese Dissertation selbst angefertigt und dabei insbesondere die Hilfe eines Promotionsberaters nicht in Anspruch genommen. Alle von mir benutzten Quellen und Hilfsmittel habe ich kenntlich gemacht und an den entsprechenden Stellen angegeben.

Axel Mayer und Norman Rose haben mich bei der technischen Durchführung und Programmierung der Analysen mit dem Softwarepaket R unentgeltlich unterstützt. Brigitte Fiege, Axel Mayer, Benjamin Nagengast, Norman Rose und Anna Zimmermann haben unentgeltlich Vorabversionen einzelner Teile des Manuskripts gelesen und mich auf Fehler und Inkonsistenzen aufmerksam gemacht. Marie-Ann Sengewald und Anna Zimmermann haben mich im Rahmen ihrer Tätigkeit als studentische Hilfskräfte im BMBF-Projekt *Faire Vergleiche* bei der Zusammenstellung des Literaturverzeichnisses entgeltlich unterstützt.

Darüber hinaus haben Dritte von mir weder unmittelbar noch mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen.

Ich habe diese Dissertation noch nicht als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht.

Ich habe weder die gleiche noch eine in wesentlichen Teilen ähnliche noch eine andere Arbeit bei einer anderen Hochschule oder Fakultät als Dissertation eingereicht.

Ich versichere, dass die oben gemachten Angaben nach meinem besten Wissen der Wahrheit entsprechen und ich nichts verschwiegen habe.

Tübingen, den _____ _____

Christiane Fiege